DOCUMENT RESUME

ED 428 945 SE 062 041

TITLE	Michigan High School Proficiency Test in Science. Tryout and Pilot Technical Report.
INSTITUTION	Michigan State Dept. of Education, Lansing.
PUB DATE	1998-01-00
NOTE	126p.
AVAILABLE FROM	Michigan Dept. of Education, MEAP Office, P.O. Box 30008, Lansing, MI 48909.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Achievement Tests; Educational Assessment; High Schools; *Science Education; *State Programs; Tables (Data); Testing Programs
IDENTIFIERS	*Michigan High School Proficiency Tests

ABSTRACT

As part of the test development process, this technical report is intended to present the technical aspects of the tryout and pilot stages of the Michigan High School Proficiency Test (HSPT) in Science. Part 1 introduces the purpose, the legislation, and the committees involved in the test development. Development of the science assessment framework and the framework structures is briefly described. Part 2 provides an overview of the exercise development of the test. Part 3 summarizes the process used in sampling, the tryout design, the rating process for constructed-response questions, reader reliability, test statistics and analyses, and other technical issues for the HSPT in science tryout and pilot administrations. Part 4 contains the summary results from student and teacher surveys conducted during the tryout stage. Relevant data tables are furnished in the appendices. (ASK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 428 945

Michigan High School Proficiency Test in Science
Tryout and Pilot Technical Report

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

3606204 1

Michigan Educational Assessment Program
Michigan Department of Education
January 1998

BEST COPY AVAILABLE

Table of Contents

	Page
List of Tables	iii
List of Figures	iv
Introduction	1
1. Evolution of the HSPT in Science	1
The Purpose of the Michigan High School Proficiency Test	1
The Expert Panel	1
Legislation Change	2
Developing the Assessment Framework to Guide the Development of the HSPT in Science	5
The Structure of the Assessment Framework	5
Committees Involved in the Development of the HSPT	6
The Technical Advisory Committee (TAC)	6
The Exercise Development Team (EDT)	6
The Content Advisory Committee (CAC)	6
The Bias Review Committee (BRC)	6
2. Exercise Development for the HSPT in Science	8
Specifications for All HSPT in Science Items	8
Specifications for Independent Multiple-Choice Items	9
Specifications for Cluster Problems	9
Developing An Integrated Science Cluster Problem	10
Specifications for the Investigation Problem	10
Specifications for the Text Criticism	10
3. HSPT in Science Tryout and Pilot	11
Sample Design and Characteristics	11
Tryout Test Design	12
Rating Process for Constructed-Response Items	14
Interrater Reliability	14
Tryout Statistics and Analyses	15
Item Difficulty	15
Test Reliability	15
Content Validity	16
Calibration Models	16
3PL/2PPC Model	16
Rasch Model	17
Calibration Analyses	18
Fit Statistics and Analyses	18
Item Discriminations	19
Equating	20
Scaling Model Selection	20
Racial and Gender Bias Analyses	21
Mantel Statistic for Ordered Response Categories	21

Standardized Mean Difference 22
Distributions of Standardized Mean Differences 23
Overall DIF Rating 23
Pilot Test 24
Pilot Sampling 25
Pilot Administration 25
General Results 25
Interrater Agreement 26
Group Descriptive Analyses 27
Gender/Ethnicity DIF Statistics 27
Summary 28

4. Student Survey and Teacher Survey 29

Science Student Survey Results 29
Conclusions from the Student Survey 30
Science Teacher Survey 30
Summary of the Teacher Survey Results 31
Overall Summary and Follow-Up 31

References 33

Appendices 34

A: Committees Assisting in the Development of the HSPT 34
Expert Panel Recommendations 39
BRC Forms 43
Michigan School Stratum Classification 45
Criteria for Writing and Editing Multiple-Choice Items 46
Checklist for Item Development 47
Checklist for Scoring Rubrics/Scoring Guide 49
B: Tryout Statistics for the HSPT in Science 50
Teacher Comment Sheet 55
C: Pilot Statistics for the HSPT in Science 57
D: Student Survey 109
Student Survey Response Means 111
Teacher Survey Results - Statements with ≥50% Schools Responding NSI 113
Teacher Survey Results - Statements with 0% Schools Responding NT 113

List of Tables

Table		Page
1	HSPT Development Timeline	4
2	Structure of the Assessment Framework for the HSPT in Science	5
3	Number of Sampled Schools in the Science Tryout by Stratum	11
4	Distribution of Students by Gender in the Tryout by Form	12
5	Distribution of Students by Ethnicity in the Tryout by Form	12
6	Configuration of the HSPT in Science	13
7	Tryout Form Composition	13
8	Interrater Agreement and Consistency Ranges for the Tryout	15
9	Raw Score Statistics by Form for the Tryout	50
10	Summary of Fit Results - 1PL/PPC for the Tryout	51
11	Summary of Fit Results - 3PL/2PPC for the Tryout	52
12	Item Flagged for Deletion Under the Fit Criteria for the Tryout	53
13	Mean and Standard Deviations of Discrimination: 3PL/2PPC	54
14	Ranges of Racial and Gender SMDs in the Tryout	23
15	Frequency Distribution of Items by Racial SMDs - Tryout	23
16	Frequency Distribution of Items by Gender SMDs - Tryout	23
17	Overall DIF Rating Classification as a Function of Gender and Race	24
18	Frequency Distribution of Items by Overall DIF Rating	24
19	Number of Students Participating in the Pilot by Form	25
20	Descriptive Statistics by Form for the Pilot	57
21	Item Statistics by Form for the Pilot	58
22	Interrater Agreement Ranges	26
23	Mean Interrater Agreement Based on First Two Readers	74
24	Interrater Agreement by Item by Form for the Pilot	75
25	Frequency Distribution for Constructed-Response Item Responses by Form	91
26	Group Descriptive Statistics	107
27	DIF Statistics (SMDs) for Gender and Ethnic Groups	108
28	Student Survey Results Summary	30
29	Teacher Survey Results Summary	31
30	Student Survey Response Means	111
31	Teacher Survey - Science: Statements with ≥ 50% Responding NSI	113
32	Teacher Survey - Science: Statements with 0% Responding NT	113

List of Figures

Figure	Page
1 - HSPT Development Process	3
2 - Configuration of Form Triplets and Quadruplets for Equating	20

Introduction

As part of the test development process, this technical report is intended to present technical information from the tryout and pilot stages of the Michigan High School Proficiency Test (HSPT) in Science. There are four major parts to this report. Part 1, Evolution of the HSPT in Science, introduces the purpose, the legislation, and the committees involved in test development. Development of the science assessment framework and the framework structures are briefly described in this part. Part 2 provides an overview of the exercise development of the test. Part 3 summarizes the process used in sampling, the tryout design, the rating process for constructed-response questions, reader reliability, test statistics and analyses, and other technical issues for the HSPT in Science tryout and pilot administrations. Summary results from student and teacher surveys conducted during the tryout stage are included in Part 4. The relevant data tables are furnished in the appendices. Operational technical reports will follow a similar format.

Part 1. Evolution of the HSPT in Science

The Purpose of the Michigan High School Proficiency Test

As required by law, The Michigan High School Proficiency Test (HSPT) was developed to provide students with an opportunity to earn state endorsement of the local diploma. Public Act 118 (P.A. 118) of 1991, Section 104(a)(subsection 7) of the School Aid Act states:

> Not later than July 31, 1993, the department shall develop and the state shall approve assessment instruments to determine pupil proficiency in communication arts, mathematics, science and other subject areas specified by the state board. The assessment instruments shall be based on the state board model core curriculum outcomes. Beginning with the graduating class of 1997, a pupil shall not receive a high school diploma unless the pupil achieves passing scores on the assessment instruments developed under this section.

The legislation initiating the development of the HSPT was introduced to respond to educators' and employers' concern that Michigan students were leaving high school without the knowledge and skills necessary to lead productive lives. Additionally, the high school diploma was awarded on the basis of local requirements. There was no consistency from school to school, nor were there, with the exception of one semester's instruction in civics, state requirements for receiving a high school diploma. The HSPT provides a consistent measure of what students should know and be able to do at the end of the tenth grade in Michigan schools.

The Expert Panel

The Expert Panel on the Michigan High School Graduation Test was convened to advise the Michigan State Board of Education on important issues surrounding the high school proficiency examination enacted by P.A. 118 of 1991. The panel consisted of national experts with first-hand knowledge and experience in large-scale testing programs (see Appendix A for list of Expert Panel members).

The Expert Panel met over three days in February and March of 1992 to examine the educational, technical, legal, fiscal and logistical issues relating to competency testing and the steps to be taken in the implementation of P.A. 118. The panel's report "Issues and Recommendations Regarding Implementation of the Michigan High School Graduation Tests" was issued in April of 1992. The

report included 51 recommendations and rationale for each of the recommendations (see Appendix A).

Legislation Change

Between the issuance of the Expert Panel Report and the development of the Assessment Frameworks for each of the content areas tested by the HSPT, new legislation was passed which dramatically changed the intent of the test. Whereas P.A. 118 had stated that the awarding and denying of high school diplomas would be determined by HSPT scores, Public Act 335 of 1993 softened the intent of the test. P.A. 335, Section 1279 states that the HSPT would be used to award state endorsements of the local high school diploma:

> Beginning with pupils scheduled to graduate in 1997, if a pupil achieves the academic outcomes required by the state board, as measured by an assessment instrument developed under subsection (8), for a state-endorsed high school diploma in 1 or more of the subject areas of communications skills, mathematics, science, and, beginning with pupils scheduled to graduate in 1999, social studies, the pupil's school district shall award a state endorsement on the pupil's diploma in each of the subject areas in which the pupil demonstrated the required proficiency. A school district shall not award a state endorsement to a pupil unless the pupil meets the applicable requirements for the endorsement, as described in this subsection. A school district may award a high school diploma to a pupil who successfully completes local district requirements established in accordance with state law for high school graduation, regardless of whether the pupil is eligible for any state endorsement... The assessment instruments shall be based on the state board model core academic curriculum outcomes...

The change in the law also changed the context in which the Expert Panel Recommendations were considered in the development of the HSPT. In addition to the Expert Panel Report, several policy decisions and subsequent policy actions shaped the development of the HSPT from the onset.

- The HSPT would align with the *Michigan Model Core Curriculum Outcomes* (State Board of Education, 1991), broad outcomes to be achieved by all students as a result of their school experiences. Fundamental to the Model Core Curriculum is the belief that the ultimate purpose of education is to permit each individual student to reach his or her optimum potential, to lead a productive and satisfying life (*The Common Goals of Michigan Education*, 1980).
- The HSPT would establish high expectations for all students.
- The HSPT would focus on the application of knowledge, problem solving and critical thinking.
- The HSPT would assess what students should know and be able to do by the end of tenth grade.
- Recognizing that what gets tested, gets taught, the HSPT would, to the extent possible in large-scale assessment, model good instructional practice.

Students earning proficient scores on the Michigan High School Proficiency Test in mathematics, science, writing and reading earn the state endorsement of the local diploma in mathematics, science and communication arts.

Table 1 and Figure 1 show the timeline and the process used by the Michigan Department of Education Michigan Educational Assessment Program (MEAP) for the development of the HSPT.

Figure 1. HSPT Development Process

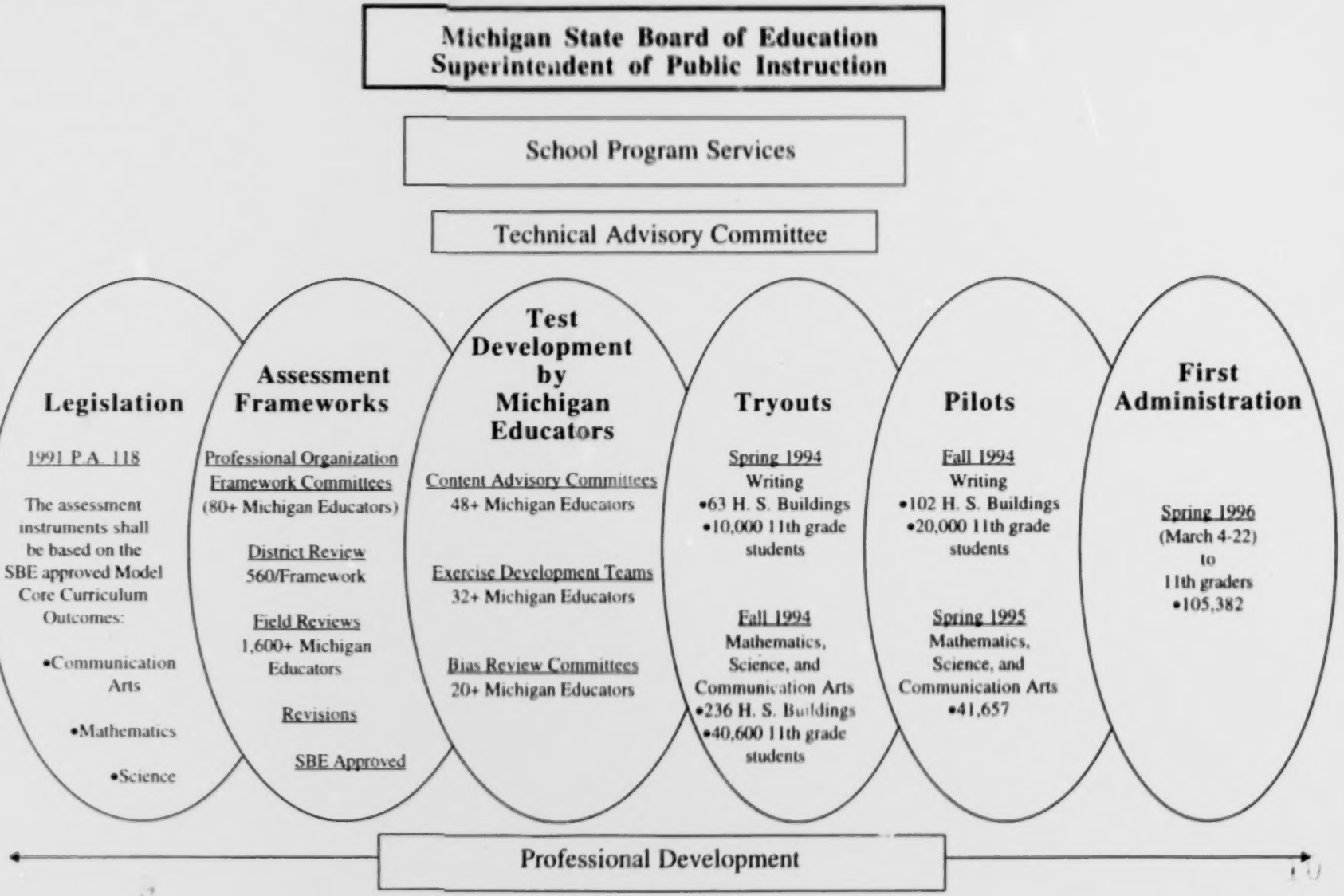


Table 1. HSPT Development Timeline

High School Proficiency Test ***Timeline 1992-1997*** Mathematics, Science, Reading, Writing	
1992-1993	**Define Test Frameworks**
November 2, 1992	Met with MRA, MSTA, MCTM and MCTE to discuss Frameworks development
January 8, 1993	Proposals to Michigan Department of Education
February, 1993	Input: Preliminary Field Review by Professional Organizations
March 31, 1993	Frameworks due to Michigan Department of Education
April 21, 1993	Michigan State Board of Education receives Frameworks
April 21 - May 31, 1993	Field Review and Comments
Summer, 1993	State Board of Education Approves Frameworks
1993, 1994, 1995	**Test Development**
Summer 1993 November 1993 January 1994	Issued RFPs Item/Exercise Development-Writing Test Item/Exercise Development-Mathematics, Science, Reading
April 1994	Tryouts-Writing Scoring, Analysis and Revision
November 1994 November 1994 April 1995	Pilots-Writing Scoring and Analysis Tryouts-Mathematics, Science, Reading Scoring, Analysis and Revision Pilots-Mathematics, Science, Reading Scoring, Analysis
1996-1997	**Test Administration Timeline**
Spring 1996	Test Administration
Fall 1996	Retest
Winter 1997	Test/Retest Award Endorsements Based Upon Results

Developing the Assessment Framework to Guide the Development of the HSPT in Science

In 1991, the Michigan State Board of Education adopted the *Michigan Essential Goals and Objectives for Science Education (MEGOSE)*. Michigan law P.A. 25 requires that the *MEGOSE* and the *Model Core Curriculum Outcomes* (1991) serve as the curriculum foundation for science education and for the HSPT in Science. The unique aspect of the *MEGOSE* (1991) and the science portion of the *Model Core Curriculum Outcomes* (1991) is that they were designed for what scientifically literate persons should know and be able to do.

The *Assessment Framework for the Michigan High School Proficiency Test in Science* (1994) was developed by Michigan Science Teachers Association (MSTA) under contract with the Michigan Department of Education. The Framework was developed by a panel consisting of science teachers, science supervisors, assessment specialists, university scientists, university science educators, a teacher certified in special education, a high school principal, and specialists from the Michigan Department of Education. A broad representation of Michigan's diverse population was involved with the project. Framework development committee members were listed in the appendix of the Framework.

The Assessment Framework was constructed to give clarity and direction to persons developing the HSPT in Science and includes detailed information on both the Core Outcomes and all of the Essential Goals and Objectives under each topic. The focus questions; the narrative sections; the related concepts, terms, and tools; and real world contexts were all provided to give a clear picture of the level of science learning that is expected from a student by the end of the tenth grade. The vocabulary and terminology used to write the assessment items were primarily taken from the Framework with any additional terminology limited to that in *Science for All Americans* (1989). The Framework includes the scientific knowledge that students are expected to learn, assessment specifications for the proficiency test, and item/exercise specifications.

On April 21, 1993, Michigan State Board of Education received the Assessment Framework developed by MSTA and authorized the Superintendent of Public Instruction to disseminate it to every school district in the state for a second round of field reviews and comments. The existing Framework represents the initial work done by MSTA with revisions based upon the field reviews and comments.

The Structure of the Assessment Framework

The Framework categories scientific literacy as using scientific knowledge, constructing scientific knowledge, and reflecting on scientific knowledge. The Using objectives are further organized into three dimensions: life science, physical science and earth science, each then was further subdivided into four or five topics. The Constructing and Reflecting objectives are not explicitly defined within the subject areas and topics as they are assessed in the context of one or more topics (see Table 2).

Table 2. Structure of the Assessment Framework for the HSPT in Science

	Life Science					Physical Science				Earth & Space Science			
Topic	Cells	Liv T	Hered	Evol	Eco	Mat	Chng	Mot	Wave	Geo	Hydro	Atm	Spce
Using Scientific Knowledge													
Constructing Scientific Knowledge													
Reflecting on Scientific Knowledge													

Under the framework design, it was projected that approximately sixty percent of the test would assess Using objectives distributed equally between earth and space, life, and physical science objectives; twenty percent would assess Constructing objectives; and, the remaining twenty percent would measure Reflecting objectives. The Constructing and Reflecting objectives would not have to be distributed equally across earth and space, life, and physical science.

Committees Involved in the Development of the High School Proficiency Test (HSPT)

The Technical Advisory Committee (TAC)

After the Expert Panel submitted its recommendations for implementing the HSPT, a subset of six core panel members was selected to form the Technical Advisory Committee (TAC) to serve in an advisory capacity during test development and implementation. Additional membership has been determined on an ad hoc basis as needed for particular expertise. The TAC has met with Michigan Educational Assessment Program (MEAP) staff periodically to provide continuous advice on technical, policy and legal issues related to the MEAP tests.

Prior to the first meeting, each TAC member received executive summaries of the assessment frameworks in mathematics, science, reading, writing, and portions of the proposal submitted by CTB/McGraw-Hill, the vendor chosen to coordinate item development for mathematics, science and reading. The TAC played an active role throughout test development and standard setting: shaping and reviewing plans, advising staff on the appropriate analyses to require of contractors and reviewing analyses provided. The TAC has been intimately involved in the program at every step and continues to be involved.

The Exercise Development Team (EDT)

The Exercise Development Team for Science was made up of nine Michigan teachers who were nominated by MDE Curriculum and MEAP staff. Members of the EDT signed a contract before item writing began. The committee members were responsible for writing all of the HSPT in Science items. All members received item writing training from CTB/McGraw-Hill. More information about exercise development for the HSPT is contained in a later section of this report.

The Content Advisory Committee (CAC)

The Content Advisory Committee for Science was responsible for the integrity of the HSPT in Science. The CAC reviewed each test item to ensure that it was appropriately related to the *Model Core Curriculum Outcomes* and the *Michigan Essential Goals and Objectives in Science Education*, as set out in the legislation. Both of these documents were approved by the State Board of Education and disseminated to school districts well in advance of the first administration of the HSPT in the spring of 1996. Items were evaluated for consistency with the criteria set out in the Assessment Framework and appropriateness for measuring proficiency in science for all students by the end of tenth grade. The CAC reviewed every test form to check for a reasonable distribution of item difficulty and for an adequate sample of the content area. Items were rejected or revised based upon decisions made by the Content Advisory Committee.

The CAC for Science was originally made up of thirteen members including high school and middle school classroom teachers, district and school science department chairpersons and college science instructors.

The Bias Review Committee (BRC)

The first Bias Review Committee was comprised of eleven members from the Michigan Department of Education and several Michigan school districts. School district personnel ranged from administrators to content area consultants to English as a Second Language (ESL) coordinators and classroom teachers. BRC members reviewed every HSPT item for possible bias

to gender, racial or ethnic groups; religious groups; socioeconomic groups; persons with disabilities; older persons; and for regional concerns. In instances where the BRC observed bias, the BRC was responsible for providing suggestions that made the test material as bias-free as possible, but did not distort or interfere with test content.

Lists of members of the above committees are in Appendix A.

Part 2. Exercise Development for the HSPT in Science

A major portion of the work in the Michigan Educational Assessment Program has been done contractually. Through the Department of Budget and Office of Purchasing, the Department of Education issues a Request for Proposals (RFP) describing the Department's testing requirements. The successful bidder must meet both quality and cost criteria as part of the evaluation process.

In order to meet the tight timeline required by legislation for development of the HSPT, CTB MacMillan/McGraw-Hill was hired to coordinate the exercise development process for the HSPT in mathematics, reading and science. CTB has years of experience in test development for national achievement tests, as well as for state assessment programs. For the HSPT, with direction from MDE Curriculum and MEAP staff, CTB provided training for the Exercise Development Team (EDT) and facilitated the EDT meetings. In addition, CTB developed the initial science item bank and test forms and ran item analyses on the tryouts and pilot tests. The CTB contract ran through the initial pilot process.

In early 1994, notebooks were sent to all committee members of the EDT to use as a resource during the development process. The notebooks, called "The Michigan Exercise Development Guidelines for Science", contained an overall schedule for exercise development and an outline of the scope of work and specific tasks for each writer. The guidelines included general item specifications and criteria for writing and editing multiple-choice and constructed-response items and for writing rubrics for the constructed-response items. The EDT completed item development by June of 1994. The following are the general item specifications used by the science EDT, as described in the Assessment Framework. Detailed item specifications for science are contained in the Exercise Development Guidelines. In addition, members of all of the content area EDTs were given the guidelines for items and rubrics that are included in Appendix A.

Specifications for All HSPT in Science Items

For this document, an "item" will be any question or task for which a response is scored. Thus, a cluster problem, for example, will include several items. All items must meet the following specifications:

1) Match with Objectives
   Each item must match one or more objectives specified in the Framework as developed from the *Michigan Essential Goals and Objectives for Science Education* (MEGOSE). The specific objective(s) should be indicated in the item description. In cases where the item matches more than one objective, one should be designated as the primary objective.

2) Real World Context
   Consistent with the definition of the objectives specified in the Framework and MEGOSE, each item must be set in the context of a real world object, event or situation. Such contexts are illustrated for each objective in column three of the charts in the framework (pp. 5-47). Other similar contexts can be used in addition to or instead of those illustrated. Real world contexts may be presented in text, pictorially or, if feasible, by video, demonstration or hands-on situation. Contexts should be such that students might experience them directly in everyday life or indirectly through popular media.

3) Subject Matter Content
   The items should be based on scientific principles rather than details and definitions. The scientific principle(s) for a given item should be explicitly expressed in the narratives or charts in the Framework and MEGOSE. In cases where this is deemed inappropriate, the

idea must be expressed in *Science for All Americans (1993)*. The scientific principle(s) for each item should be identified in the item description with the source cited.

4) Technical Vocabulary

The technical vocabulary used in items and required in item responses should be based on the terms included in the narrative or column two of the charts in the proficiency test framework and MEGOSE. In cases where this is judged to be too restrictive, such terms must be found in appropriate sections of *Science for All Americans* (1989). Such cases should be identified in the item description.

5) Freedom From Bias

Real world contexts and non-technical vocabulary for items should be familiar to all students. Items should provide balance where contexts or vocabulary might be more familiar to one group than another (e.g., different genders, geographical regions, or ethnic or cultural backgrounds). Items and tests must be reviewed for bias by appropriate expert panel(s).

6) Readability

Our goal for readability is that students have access to the information on the test and be able to use it without hindrance of reading ability. Exclusive of technical vocabulary, the readability of items should be at or below that adopted for the entire HSPT.

7) Scientific Accuracy

The designated or example correct response for items must be scientifically accurate. Scientific accuracy must be certified by a content expert panel(s), taking into account the necessity to express scientific ideas in terms of the restricted technical vocabulary.

8) Special Education

All students seeking state endorsement on Science Proficiency will demonstrate their proficiency by performance on the HSPT. Students with special needs should be offered the opportunity to achieve science proficiency as are their peers. Accommodations for administering and responding to the proficiency test should be made to compensate for and/or address disabilities of such students.

Specifications for Independent Multiple-Choice Items

Independent items present a brief description of a real world context and pose a single question about it. Each item assesses one core outcome. The purpose of these items is to test a wide designated sample of outcomes.

Specifications for Cluster Problems

A cluster problem presents a real world context (an event, a situation or an object) and asks a series of questions about it. The proficiency test will include a series of such problems, including one for each of the three dimensions of life science, physical science, and earth and space science. A cluster problem will include four multiple-choice questions and one constructed-response question, which will involve all three kinds of activities defined in the Framework (using, constructing and reflecting on scientific knowledge).

Developing an Integrated Science Cluster Problem

An integrated science cluster problem addresses objectives from at least two of the three disciplines (life, earth and space, physical). It should address the reflecting and constructing objectives listed under the cluster specification, but should also include at least one of the following:

R9 Describe the advantages and risks of new technologies
R15 Evaluate alternative long-range plans for resource use and by-product disposal

Response formats should follow the general recommendations for cluster problems.

The integrated science cluster problem requires that a scenario is written which uses a real life setting, and whose solution requires that students use knowledge from two or more disciplines. The MEGOSE provides examples of how each of the essential goals and objectives is referenced to a connecting theme of science.

Specifications for the Investigation Problem

The investigation portion of the HSPT in Science requires students to read a report of an experiment conducted by students and asks them to respond to two or three constructed-response questions about the report that cover outcomes of constructing scientific knowledge only. The subject matter topic must be different from that of the text criticism. The investigation problem focuses on experimental design.

Specifications for the Text Criticism

The text criticism presents students with a passage to read from the popular press (newspaper or periodical) and requires them to respond to two or three constructed-response questions covering only Reflecting core outcomes. The questions require students to gather and synthesize information from the passage and to evaluate the strengths and weaknesses of claims, arguments or data included in the passage. The questions may also require students to describe some limitations of scientific knowledge relevant to the passage.

Part 3. HSPT in Science Tryout and Pilot

After the Exercise Development Teams completed items for each content area to be tested on the HSPT, the Content Advisory Committees and the Bias Review committee reviewed all items. Tryouts were scheduled for the items that survived this initial committee review. Statistical data from tryouts and pilots are part of the information used to determine which items merit further consideration for use on "live" or operational tests. In addition, participating teachers are asked to return comment sheets describing problems with the directions and/or items and noting administration details, such as the amount of time it took the majority of students to complete the test (see Appendix B for a sample). Comments from teachers are particularly helpful in making decisions about items and test forms.

Sample Design and Characteristics

Data for the HSPT in Science tryout and pilot were collected using the same procedures. To ensure representativeness, cluster sampling combined with stratification was used to sample from Michigan public schools. Michigan schools are classified into seven strata by resident population size of the community where the school is located (see Appendix A for stratum classifications). Schools participating in the tryout were randomly sampled from each stratum roughly proportional to the population proportions. The number of sampled schools in the science tryout by stratum is listed in Table 3 below.

Table 3. Number of Sampled Schools in the Science Tryout by Stratum

Stratum	# of Schools Sampled	Total # of Schools in the Stratum	% of Stratum
1	5	49	10.2%
2	7	64	10.9%
3	12	106	11.3%
4	8	62	12.9%
5	2	7	14.3%
6	23	232	9.9%
7	19	218	8.7%
undefined[1]	5	NA	NA
Total	80	738	--

The sampled schools were considered representative of Michigan student population in gender, ethnicity, and school size. Distributions by gender and ethnic groups for the science tryout by test form are shown in Tables 4 and 5.

Schools participating in the tryout were not sampled again for the pilot. Schools that were sampled for the tryout or pilot but did not participate were replaced by schools with similar characteristics to keep the representativeness of the sample. Also, schools participating in the science tryout or pilot were not selected in the reading or mathematics tryouts and pilots.

[1] These schools were either alternative or adult education schools.

Table 4. Distribution of Students by Gender in the Tryout by Form

Form	Total # of Students Tested	# of Males	# of Females
20	1044	523	521
21	1058	525	533
22	1056	532	524
23	1073	518	555
24	1033	527	506
25	1036	495	541
26	902	424	478
27	955	428	527
28	964	436	528
29	953	470	483
Total	10074	4878	5196

Table 5. Distribution of Students by Ethnicity in the Tryout by Form

Form	# of Students Tested	Am. Indian N (%)	Asian N (%)	Black N (%)	Hispanic N (%)	White N (%)	Multi-Racial N (%)	Other N (%)
20	1044	9 (0.9)	17 (1.6)	93 (8.9)	42 (4.0)	804 (77.0)	44 (4.2)	35 (3.4)
21	1058	17 (1.6)	11 (1.0)	82 (7.8)	47 (4.4)	813 (76.8)	47 (4.4)	41 (3.8)
22	1056	18 (1.7)	17 (1.6)	161 (15.2)	38 (3.6)	722 (68.4)	45 (4.3)	55 (5.2)
23	1073	25 (2.3)	14 (1.3)	163 (15.2)	17 (1.6)	786 (73.3)	30 (2.8)	38 (3.6)
24	1033	13 (1.3)	30 (3.9)	193 (18.7)	20 (1.9)	694 (67.2)	47 (4.5)	36 (3.5)
25	1036	9 (0.9)	33 (3.2)	176 (17.0)	20 (1.9)	736 (71.0)	30 (2.9)	32 (3.1)
26	902	15 (1.7)	41 (4.5)	164 (18.2)	14 (1.6)	613 (68.0)	28 (3.1)	27 (2.9)
27	955	17 (1.8)	21 (2.2)	128 (13.4)	18 (1.9)	690 (72.3)	30 (3.1)	51 (5.3)
28	964	18 (1.9)	16 (1.7)	130 (13.5)	13 (1.3)	716 (74.3)	26 (2.7)	45 (4.7)
29	953	16 (1.7)	18 (1.9)	73 (7.7)	35 (3.7)	728 (76.4)	31 (3.3)	52 (5.5)
Total	10074	157 (1.6)	218 (2.2)	1363 (13.5)	264 (2.6)	7302 (72.5)	358 (3.6)	412 (4.1)

Tryout Test Design

There were 10 tryout forms in science, configured as shown in Table 6 below. Each of the forms contained four item formats: independent multiple-choice questions; cluster problems; text criticism; and investigation.

Table 6. Configuration of the HSPT in Science*

Science Subject Area	Life Science			Physical Science			Earth Science			Integrated Science			Number of Items
Objective Category	U	C	R	U	C	R	U	C	R	U	C	R	
Cluster Problems (4 MC items and 1 constructed-response item each)	3	1	1	3	1	1	3	1	1	3	1	1	20
Independent Multiple-Choice Items	7	2	1	7	2	1	7	2	1				30

Science Subject Area	Life, Physical, or Earth Science		Number of Problems
Objective Category	C	R	(2 items each)
Text Criticism		1	1
Investigation	1		1

* Legend: Activities
U - Using scientific knowledge - about 60% of items
C - Constructing scientific knowledge - about 20% of items
R - Reflecting on scientific knowledge - about 20% of items

Each form contained 30 independent multiple-choice questions, 10 for each subject area. There were also four cluster problems per form, one per subject area and a fourth that integrated at least two of the three subject areas. Each cluster problem was comprised of four multiple-choice questions and one constructed-response question. In addition to the independent multiple-choice questions and the cluster problems, there was one text criticism and one investigation problem on each tryout form. The text criticism contained two constructed-response questions covering only Reflecting Core Outcomes. The investigation problem consisted of two constructed-response questions that covered only Constructing Core Outcomes and from a different subject area other than the text criticism problem.

The Science tryout involved 10,074 students in grade 11 during the late fall of 1994. Each student took one tryout form. Since there were 10 forms and no items overlapped between any two forms, randomly equivalent group equating was used. To avoid exposing all forms to a participating school, forms were divided into four groups of triplets and two groups of quadruplets, related by theme (Table 7). A school was randomly assigned to take only one group of forms. The forms within each triplet (or quadruplet) were then spiraled and administered to students within a classroom so that no students sitting next to each other would have the same form. This design permitted the equating of forms between triplets (or quadruplets) through the assumption of randomly equivalent groups of different participating schools taking the same form, but in different compositions. Forms in different triplets or quadruplets were equated by the Stocking and Lord (1983) procedure applied to the items in the common form. Additional information about equating will be presented in a later section of this report.

Table 7. Tryout Form Composition

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Form 20	Form 22	Form 24	Form 26	Form 29	Form 23
Form 21	Form 23	Form 25	Form 27	Form 20	Form 25
Form 22	Form 24	Form 26	Form 28	Form 21	Form 27
			Form 29		Form 28

BEST COPY AVAILABLE

Rating Process for Constructed-Response Items

All multiple-choice items were machine-scored. All constructed-response items were hand-scored by two readers. Readers were trained to implement the Michigan scoring guides. A number of quality control procedures were taken to ensure interrater reliability. Sets of actual student papers was used as anchor papers to illustrate responses exemplifying each of the possible score points for a response. Student responses were also used in check sets throughout the scoring process to ensure that readers were consistently applying Michigan standards. Table leaders conducted "read-behinds" by re-scoring sets of student responses to check the consistency of readers at their tables. For all constructed-response items, if the two readers disagreed by more than one point, a third reader was asked to adjudicate the scores. This situation rarely occurred. If two readings were sufficient, the item score was the sum of the two readings. If three readings were required, the item score was the sum of the three readings multiplied by 2/3, and rounded to the nearest integer. This process provided constructed-response items with 3, 5, or 7 score levels in science.

Interrater Reliability

Indices of interrater reliability, in the form of ranges of exact agreement and consistency, are presented by form in Table 8 below. For this analysis, the agreement is defined as the percent of times that the first reader agreed, within one point, with the second reader on the common items read by both readers:

$$\text{Agreement} = \frac{\text{\# of Items Reader 1 within One Point of Reader 2}}{\text{\# of Common Items Read by Readers 1 and 2}} \times 100 \qquad (1)$$

The agreement range describes the lowest and highest agreement rates seen among all readers. Consistency is defined as the percent of times the first reader agreed, within one point, with the second *or* the third reader:

$$\text{Consistency} = \frac{\text{\# of Items Reader 1 within One Point of Reader 2 or Reader 3}}{\text{\# of Common Items Read by Readers 1 and 2}} \times 100 \qquad (2)$$

The consistency range spans the readers who had the smallest and largest consistency rates. Consistency rates must be at least as large as agreement rates.

Both agreement and consistency ranges were generally small for the HSPT in Science tryout, with upper bounds that were often at 100%. Only one form in science, Form 20, had an agreement range that dipped below 98%, due to one reader who completed only 8 total readings, compared to an average of hundreds of readings for the remaining readers.

Table 8. Interrater Agreement and Consistency Ranges for the Tryout

FORM NUMBER	AGREEMENT RANGE	CONSISTENCY RANGE
20	88 - 100%	88 - 100%
21	99 - 100	99 - 100
22	98 - 100	98 - 100
23	99 - 100	99 - 100
24	98 - 100	99 - 100
25	98 - 99	99 - 100
26	99 - 100	99 - 100
27	99 - 100	99 - 100
28	99 - 100	99 - 100
29	100	100

Tryout Statistics and Analyses[2]

Item Difficulty

Ranges of item difficulty (p-values) and item test correlations are presented in Table 9 (Appendix B). Rather than presenting the full range, which usually is not very informative because of the occurrence of outliers, the statistics are presented for the center 80 percent of the items in each form. That is, the items were rank-ordered in terms of p-values, and the values tabled for items at the 10th and 90th percentiles. For example, if a test had 40 items, p-values for the 4th and 36th most difficult items would be tabled. These ranges of p-values indicate that there was a good spread of item difficulties. Although not presented in this table, other analyses indicated that the constructed-response items tended to be among the more difficult items in each form.

The "Collapsed Levels" columns in Table 9 indicate items where there were too few examinees who scored in a particular level so that scaling of that level for that item could not take place. In general, if there were fewer than 4 students with scores in a level, calibration could not occur. When calibration cannot occur, adjacent levels are collapsed. There were few levels for few items in which collapsing was necessary. The sparse levels tended to be those for the highest score levels of the most difficult items. While collapsing of levels can be important in a final operational calibration, collapsing of levels has little impact in a tryout.

The average percentage of maximum score (%MS) ranged from 40 to 52 for all 10 tryout forms. Thus, the test was fairly difficult for these students, but not so difficult as to create floor effects.

A final check was performed after the initial item analyses identified items that were very difficult or had low item-test correlations. No science items proved to be problematical under this consideration. Three science items (in two forms) were flagged for multiple correct answers; these items were not scored in any further analyses.

Table 9 in Appendix B contains raw score statistics for the forms.

Test Reliability

The reliability of a test indicates how well the test items "hang together." For the HSPT, reliability values are determined using internal consistency formulas, which indicate that the tests are measuring the same thing (within a particular test), and that students are answering consistently.

[2] See Appendix B for tryout statistics.

Cronbach's alpha is used when there is a combination of multiple-choice and constructed-response items.

The coefficient alpha reliabilities were reasonable for the number of items in the science tryout, ranging from .86 to .91 (Table 9). Alpha coefficients were computed in two ways, both including all items and excluding each individual item in each form of the tryout. The two outcomes were not proven to be significantly different.

Content Validity

As stated earlier, the current assessment is based on the *MEGOSE* (1991). Because the current test is an achievement test used to endorse individual students' diplomas in science, the most important type of validity to assess is content validity. To verify content validity, the test items must match the specified objectives given in the test blueprint or assessment framework.

Like all published achievement tests, the HSPT in Science has a blueprint which indicates the objectives to be tested (see Table 2 earlier). Not all objectives are tested in any given form of a test. Both easy and hard items are used in every form of the test to balance the difficulty level of the items and to equate the different versions of the test to one another. The sample of items chosen for a version of the test represents the domain of all possible test items that fit the blueprint. For a student to do well on the test, he or she must have mastered the entire domain, not just bits and pieces.

It was stated earlier in this report that the EDT in Science wrote all the tryout items based on the science blueprint and framework documents. The CAC verified that each test question meets the objective it is supposed to measure and fits the blueprint or framework. The BRC verified that the items are not disadvantaging any particular group.

Calibration Models

According to item response theory, item parameters are relatively invariant to changes in examinee groups. The important practical impact of this property is that the parameters of large numbers of items can be estimated even though each item is not answered by every examinee. This is known as person-free item calibration. The purpose of calibration is to estimate item parameters (e.g. item difficulty) as accurately as possible.

There are many calibration models. For the development of the HSPT, all calibration analyses were replicated using two sets of models, as recommended by the Technical Advisory Committee: (1) a combination of three-parameter logistic and two-parameter partial-credit models (3PL/2PPC) and (2) a combination of Rasch logistic and Rasch partial-credit models. The logistic models were used to analyze multiple-choice items and the partial-credit models were used to analyze constructed-response items. The purpose was to compare which set would more appropriately reflect the data.

<u>3PL/2PPC Models</u>

The three-parameter logistic (3PL) model (Lord, 1980) allows items to vary in difficulty and discrimination and non-zero lower asymptotes ("guessing values"). It is commonly applied to multiple-choice items in tests like the HSPT, where guessing of correct answers can occur.

$$P_j(\theta) = P(X_j = 1 | \theta) = c_j + \frac{1 - c_j}{1 + \exp[-1.7a_j(\theta - b)]} \qquad (3)$$

where θ = examinee's latent trait
a_j = item discrimination parameter for item j
b_j = difficulty parameter for item j
c_j = guessing parameter for item j
X_j = observed score for item j
$P_j(\theta)$ = probability of answering item j correctly given person ability θ

For the jth open-ended item with m_j levels, the item scores were integers ranging from 0 to m_j - 1 levels. A two-parameter partial-credit (2PPC) model allows items to vary in both difficulty and discrimination. It was used to calibrate constructed-response items (Yen, 1993). This model can be seen as a special case of Bock's (1972) nominal model and is the same as Muraki's (1992) "generalized partial-credit model," which is used with the National Assessment of Educational Progress (NAEP) test. The probability of a student with ability θ having a score at the kth level of the jth item is

$$P_{jk}(\theta) = P(X_j = k-1 \mid \theta) = \frac{\exp(z_{jk})}{\sum_{i=1}^{m_j} \exp(z_{ji})}, \qquad k = 1, \ldots, m_j \tag{4}$$

where

$$z_{jk} = \alpha_j (k-1)\theta - \sum_{i=0}^{k-1} \sigma_{ji} \qquad i = 1, \ldots, k, \ldots, m_j \tag{5}$$

and

$$\sigma_{j0} \equiv 0.$$

α_j is the item discrimination. σ_{ji} is related to the difficulty of the item levels: the trace lines for adjacent scores levels intersect at σ_{ji} / α_j.

The 2PPC model is as follows:

$$P_{j2}(\theta) = P(X_j = 1 \mid \theta) = \frac{1}{1 + \exp[-\alpha_j \theta + \sigma_{j1}]}. \tag{6}$$

Then,

$$a_j = \alpha_j / 1.7, \tag{7}$$

$$b_j = \sigma_{j1} / \alpha_j; \tag{8}$$

Conversely,

$\alpha_j = 1.7a_j$ and $\sigma_{j1} = 1.7a_j b_j$.

Rasch Models

The Rasch logistic model was used for multiple-choice items. This model allows items to vary in terms of difficulty, but all items were assumed to have the same discrimination (1.0) and a zero asymptote:

$$P_j(\theta) = P(X_j = 1 | \theta) = \frac{1}{1 + \exp[b_j - \theta]}. \quad (9)$$

Because of these simplified assumptions, for a two-level item,
$a_j = \alpha_j = 1$,

$b_j = \sigma_{j1}$.

Masters' (1982) Partial Credit model was used for the constructed-response items. In formula,

$$P_{njx} = \frac{\exp \sum_{i=0}^{x} (\theta_n - b_{ji})}{\sum_{k=0}^{m} \exp \sum_{i=0}^{k} (\theta_n - b_{ji})}, \quad x = 0, 1, 2, ..., m_j \quad (10)$$

where P_{njx} is the probability of person n scoring x on constructed-response item j.

Calibration Analyses

Item parameters and θ estimation were conducted using the CTB-owned program PARDUX (Burket, 1991; 1995) and the commercial software BIGSTEPS (Linacre & Wright, 1993). PARDUX employs a marginal maximum likelihood procedure, implemented with an EM algorithm. Evaluations of the accuracy of the program with real and simulated data (Fitzpatrick, 1994) have found it to be at least as accurate as the Rasch program BIGSTEPS. The MEAP office traditionally uses BIGSTEPS.

For comparison purposes, BIGSTEPS estimates using the Rasch model were obtained in addition to the PARDUX analyses for Mathematics Form 14 in Group 6 and Reading Form 1 in Group 1. The correlations between parameters obtained by the two programs were 1.00. The two programs produced very similar estimates, with the estimates being the most similar for the item score levels where the most data were available.

Fit Statistics and Analyses

Item fit was evaluated from PARDUX with a statistic comparing observed and predicted trace lines. This fit statistic is a generalization of the Q_1 statistic described by Yen (1981). Standardized fit values, referred to as Z statistics, can be compared over items and models. In addition, observed and predicted trace lines were compared graphically.

Rules of thumb were developed for flagging items for misfit. Recall that each item was scaled in two different samples. An item was flagged if it met either of the following criteria:

(1) Zs ≥ 4.0 in both samples, or

(2) (one Z ≥ 4.0) and (4.0 > the other Z ≥ 3.0), and a plot of expected and observed trace lines failed to demonstrate reasonable fit. (Note: A Z score is a standardized item fit score with a mean of zero and a standard deviation of 1.)

These rules of thumb for flagging misfit items can be compared in terms of stringency to the criterion used by CTB/McGraw-Hill for the tryout of multiple-choice items for major achievement batteries, such as the California Achievement Tests, and the Comprehensive Tests of Basic Skills. For those tests, Zs of 4.6 are flagged, even though their sample sizes are usually at least twice the size of ones used in the present study. As sample size increases, the power of the fit statistic increases. Thus, the flagging criteria used in this study are less stringent than used by CTB/McGraw-Hill in some other testing programs.

Summaries of item fit results are presented in Tables 10, 11, and 12 (Appendix B). More items from the Rasch model had large Z values and were flagged for misfit than those from the 3PL/2PPC model. With the Rasch model, 7.5% (6/80) of the constructed-response items were flagged to be misfit, while with the 3PL/2PPC model, only one constructed-response item showed misfit. However, for the 3PL/2PPC model, there were items whose parameters could not be estimated, called non-convergent items. These items were often difficult items with low discrimination values. For the Rasch model, on the other hand, parameter estimates were convergent for all items. Thus, neither model effectively described an item performance when its observed trace line was essentially flat and had weak relationship to the predicted trace line. It should be noted that all the results shown here are from the software program PARDUX. Verification of the results from the software BIGSTEPS, which was designed specifically for Rasch model analysis, showed that some items that were misfit with the PARDUX were proved to be fit with BIGSTEPS.

Item Discriminations

The item discriminations (Table 13, Appendix B) were systematically lower for the constructed-response items than for the multiple-choice items. On the average, the constructed-response items had discriminations that were 47% of the values for the multiple-choice items for science. Discriminations reflect how sharply performance can be categorized into successive score levels. It is not surprising that this categorization is less distinct with items that involved human evaluations of multiple levels of complex student performance.

The fact that the constructed-response items had lower discriminations does not mean that these items are "less important" or contribute less information to the overall test score. The formula for item information is the following:

$$I(X_j \mid \theta) = a_j^2 \sigma^2(X_j \mid \theta) \qquad (11)$$

The item information is a function of both the item discrimination (a_j^2) and the variance of the item scores (σ^2). Items with more score levels tend to have substantially greater score variances, thus adding to the information they provide. Despite their lower discriminations, the constructed-response items provided substantial amounts of information. Under the Rasch model, where all items are assumed to have the same discrimination, items with more score levels must be described as providing more information.

Table 13 (Appendix B) presents means and standard deviations of discrimination parameter estimates for all forms.

Equating

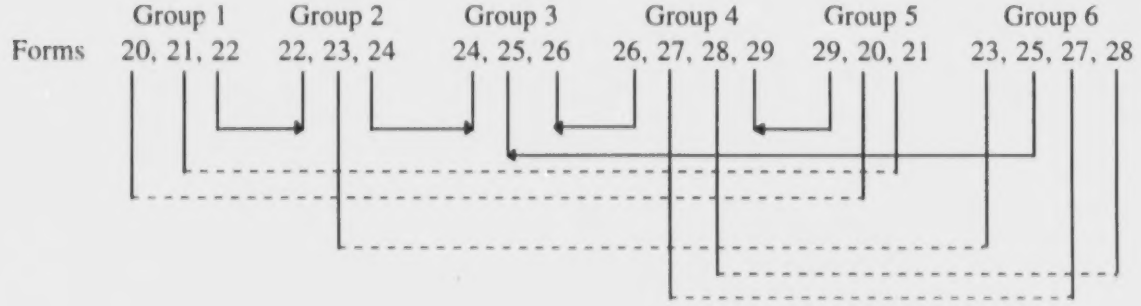
Test equating is necessary whenever one of two situations below occurs:

1. The tests are at comparable levels of difficulty and the ability distributions of the examinees taking the tests are similar. This is called "horizontal equating."
2. The tests are at different levels of difficulty and the ability distributions of the examinees are different. This is called "vertical equating."

For the HSPT tryouts, horizontal equating was used because multiple forms were developed for each subject area and administered to randomly equivalant groups. The purpose of equating is to transform the scores of examinees taking form *X* to equivalent scores in form *Y* so that these scores can be compared to the scores of examinees taking form *Y*.

The equating process was conducted for both the Rasch and the 3PL/2PPC models here. The within-triplet theta (or scale score) distributions were aligned. The Stocking and Lord (1983) procedure was applied to the forms in common to the triplets or quadruplets (Forms 22, 24, 25, 26, and 29), as indicated by the solid lines in Figure 2.

Figure 2. Configuration of Form Triplets and Quadruplets for Equating



The dotted lines indicate forms that were not included in the Stocking and Lord links (Forms 20, 21, 23, 27, 28). These forms, therefore, could be used as a check on the adequacy of the equating. Forms 20 and 21 were of particular importance because the parameters from groups 1 and 5 were the "furthest apart" in terms of the linkings; that is, five Stocking and Lord links and five equivalent group links tied them together. By comparing the Form 20 test characteristic function based on the parameters from Group 1 to that based on Group 5, the adequacy of the link network could be double-checked. Similar checks could be done for forms 21, 23, 27 and 28. The checks showed that both models produced good equating results.

Scaling Model Selection

The advantages of using a Rasch model are its simplicity and elegance. Also, if data are scarce, Rasch model predictions tend to be more stable than those from a model with more parameters. The disadvantage of the Rasch model is that its simplifying assumptions may be inappropriate for a particular data set. The major advantage of the 3PL model is its less restrictive assumptions that

permit more accurate description of data. The major disadvantage of the model is that it requires a large number of examinees to provide sufficient data for parameter estimation. However, this was not a problem for the HSPT tryouts.

For the HSPT tryout data, Rasch models provided more misfit items (particularly for constructed-response items) than the 3PL/2PPC models did, but the Rasch models did provide parameter estimates for all items. The 3PL/2PPC models produced better item estimates for most items but failed to converge for some other items in calibration (i.e., no estimates for those items). The TAC recommended the use of Rasch models over the 3PL/2PPC models for a large-scale assessment such as the HSPT, based on the empirical evidence and other technical considerations.

Racial and Gender Bias Analyses

Mantel Statistic for Ordered Response Categories

A Mantel-Haenszel methodology was used in the evaluation of the tryout items for differential item functioning (DIF). A statistic proposed by Mantel (1963) was obtained for specified racial and gender groups:

$$\chi^2 = \left[\sum F_k - \sum E(F_k)\right]^2 / \sum Var(F_k), \qquad (12)$$

where F_k, the sum of scores for the focus group at the kth level of the matching variable is:

$$F_k = \sum y_j n_{Fjk}, \qquad (13)$$

Readers are referred to Zwick et al. (1993) for a description of the terms of the statistic. The Mantel statistic, while necessary for the assessment of DIF in the constructed-response items in each of the three content areas, reduces to the Mantel-Haenszel chi-square statistic (without continuity correction) when applied to the multiple-choice items. The Mantel statistic explicitly takes into account the possible ordering of the categories of the polytomous items, as opposed to a procedure proposed by Mantel and Haenszel (1959) that provides for a comparison of the reference and focus groups with respect to their entire response distributions. The Mantel statistic has a chi-square distribution with one degree of freedom.

Because the number of students in the minority groups taking each form was relatively small (almost always less than 200 per form) and the number of levels for some of the constructed-response items was large (greater than five), when item scores were obtained by summing judges' ratings, the number of levels was collapsed for some constructed-response items. After collapsing adjacent levels, the number of remaining levels that were evaluated for each constructed-response item was half the maximum number of points plus one, or the same number of levels specified by the scoring rubrics for each item for each individual reader.

As specified by MDE for a sample of schools that were supplied to CTB/McGraw-Hill, item responses were analyzed for gender bias by evaluating DIF against females (focus group), with males as the reference group. The number of females in these analyses was large, approximately half of the roughly 1000 students who took each form.

The particular racial groups that were evaluated in the racial bias analyses were determined by the numbers of students in these groups that took the 29 tryout forms in the three content areas. The only group, excluding whites, that had appreciable numbers taking each form was African-Americans. Seventeen of the forms were administered to more than 100 African-Americans. The 12 forms that had fewer than 100 African-Americans were due to two schools with large African

American enrollments dropping out of the sample and the failure to receive scores from a third school. A fourth school did not have as large an African-American population as expected.

After African-Americans, no defined racial group had consistently as many as 30 students taking each form. Consequently, Mantel statistics were obtained for a single (focus) racial group, African-Americans, treating whites as the reference group in the racial bias analysis.

Mantel racial and gender statistics were obtained for each form of the science test by stratifying on total score. A total of 62 out of 540 science items had a Mantel statistic that indicated racial DIF at a .05 significance level compared to 176 items that were flagged at the same significance level for gender DIF. Standardized mean differences were employed to provide further investigation of item bias.

Standardized Mean Difference

Although the number of items that had significant Mantel gender statistics in each of the three content areas is substantially larger than the number of items having significant racial statistics, there are three reasons why the number of significant statistics cannot be considered to reflect the magnitude of DIF within each content area. First, the Mantel statistic is *asymptotically* distributed as chi-square, requiring a minimum expected number of five students within each of the cells defined by the combinations of strata and item levels. For the racial analysis, this assumption is frequently violated.

Second, a significant Mantel statistic rejects the null hypothesis of no DIF against the alternative hypothesis of DIF either against the focus *or* the reference group. Hence the number of significant Mantel statistics does not reflect solely DIF against the assessed focus group.

Finally, the much larger sample sizes for the female focus group relative to the African-American focus group results in more statistically powerful tests (i.e., tests that are more capable of correctly rejecting the null hypothesis of no DIF) in the gender analysis. The Mantel statistics for gender can detect the presence of smaller, and perhaps practically insignificant, amounts of DIF than the corresponding statistics from the racial analysis. An analysis of DIF that is more suitable to demarcating practically significant amounts of DIF across both racial and gender analyses would utilize an effect size index.

Unfortunately, while an effect size index in the form of the Mantel-Haenszel common odds ratio estimate, alpha, is available for the dichotomously scored items, no *single* analogous odds ratio-estimate is available for the polytomous items. The standardized mean difference (SMD) noted by Zwick et al, (1993) offers an acceptable alternative.

$$SMD = \sum p_{Fk} m_{Fk} - \sum p_{Fk} m_{Rk}, \qquad (14)$$

where $p_{Fk} = n_{F+k}/n_{F++}$ is the proportion of focus group members who are at the kth level of the matching variable, $m_{Fk} = (1/n_{F+k})\ (\sum y_t n_{Rtk})$ is the mean item score for the focus group at the kth level, and $m_{Rk} = (1/n_{R+k})\ (\sum y_t n_{Rtk})$ is the analogous value for the reference group. As an effect size index, the SMD statistic takes into account the natural ordering of the response levels of the items and has the desirable property of being based only on those ability levels where members of the focus group are present. A positive value for a SMD reflects DIF in favor of the focus group, while a negative SMD indicates DIF in favor of the reference group.

Distributions of Standardized Mean Differences

Both racial and gender SMDs were obtained for the items in every form and are presented with the Mantel statistics. Ranges of the racial and gender SMDs for the science tryout are shown below:

Table 14. Ranges of Racial and Gender SMDs in the Tryout

Content Area	Racial	Gender
Science	- .32 to .24	- .25 to .24

An evaluation of both the Mantel and the SMD statistics for the racial comparisons suggested that levels of standardized mean differences that have practical significance could be determined. Statistically significant (p = .05) racial Mantel statistics were often associated with SMDs that had absolute values of .10 and greater. Setting a criterion of -.10 for a determination of practically significant DIF, representing a one tenth of a score point decrement in focus group performance relative to the reference group (controlling for ability), would allow a goal of limiting the conditional between-focus-and-reference-group difference to no more than one score point in any form. The distribution of SMDs for science below appears to permit the construction of forms having 10 or fewer items demonstrating DIF against either a racial or gender group that an individual form could have and still attain the maximum one score point conditional group difference goal. A maximum of one score point difference is desirable, given the high-stakes nature of the test.

Table 15. Frequency Distribution of Items by Racial SMDs - Tryout

(SMD≤-.30)	(SMD≤-.20)	(-.19≤SMD≤-.10)	(-.09≤SMD≤.09)	(.10≤SMD≤.19)	(SMD≥.20)	(SMD≥.30)
1 items	5 items	41 items	427 items	65 items	2 items	0 items

Table 16. Frequency Distribution of Items by Gender SMDs - Tryout

(SMD≤-.30)	(SMD≤-.20)	(-.19≤SMD≤-.10)	(-.09≤SMD≤.09)	(.10≤SMD≤.19)	(SMD≥.20)	(SMD≥.30)
0 items	6 items	41 items	454 items	36 items	3 items	0 items

Overall DIF Rating

The distribution of racial and gender SMDs under the criterion of -.10 for practically significant DIF allows the construction of an overall rating of DIF that combines both racial and gender DIF against the focus groups. An overall rating is a useful index in the development of the pilot or operational forms. Content editors can utilize test development software to select items in a manner that minimizes DIF against both focus groups.

A useful overall index of DIF might allow several gradations of the practical severity of both racial and gender DIF. An item could be considered to manifest a lower degree of practically significant DIF against a racial or gender group if the SMD ranged between -.10 and -.19 and a more serious degree of DIF if the SMD was less than or equal to -.20. An item would accumulate one point on the overall rating scale if the racial SMD fell in the former category and two points if the racial

SMD fell in the latter category. Similarly, an item would accumulate an additional point on the overall scale if the gender SMD fell in the former category and two points if in the latter. Consequently, if an item demonstrates neither of the two levels of practically significant racial DIF and neither of the two levels of practically significant gender DIF, the item's overall rating would be one (zero would seem to be a less desirable alternative because it connotes the absence of DIF). An item would obtain the maximum overall rating of five if both racial and gender DIF was of the more serious kind. An overall rating of two would imply the item had a racial or gender SMD between -.10 and -.19, but not both. An overall two, three, or four could be obtained by various combinations of lower and higher levels of practically significant racial and gender DIF. All possible overall ratings are described in the table below.

Table 17. Overall DIF Rating Classification as a Function of Gender and Race

		Race DIF	
Gender DIF	(.09 ≥ SMD ≥ -.09)	(-.10 ≥ SMD ≥ -.19)	(-.20 ≥ SMD)
(.09 ≥ SMD ≥ -.09)	1	2	3
(-.10 ≥ SMD ≥ -.19)	2	3	4
(-.20 ≥ SMD)	3	4	5

Table 18. Frequency Distribution of Items by Overall DIF Rating

DIF Rating	1	2	3	4	5
# of items	457	67	13	2	1

Items with a DIF rating of two or higher were subject to an additional review by the Bias Review Committee and the Content Advisory Committee for any apparent bias. If none was found and the item was considered to adequately measure the test content, it was kept.

Pilot Test

Items that survive the tryout stage are then piloted before they are used in an operational test. Frequently, 25-50% of items tried out are discarded at the tryout stage. Based on review of the tryout results, CTB worked with the CAC for Science and MDE staff to refine items and scoring rubrics before piloting began. Sufficient numbers of items survived the tryout to construct eight pilot forms of the test. A major change was that one multiple-choice item was eliminated from all clusters in all forms, leaving three multiple-choice and one constructed-response item for a cluster, the number that remained for the operational tests.

The purposes of the pilot administration were to:

- check if revisions based on the tryouts were successful, or whether an item should never be used;
- produce 6 equivalent forms of the High School Proficiency Test in Science that could be used interchangeably in future administrations;
- examine characteristics of the revised items in each form; and,
- examine technical soundness of the reconstituted forms for operational administrations.

CTB made all necessary revisions of the assessment materials suggested by the CAC and MDE. They also prepared the test booklets, answer documents, administration manuals and all supporting materials for the pilot administration.

Pilot Sampling

As in the tryout, the target population for the pilot was all eleventh graders in Michigan, including students in both public and private schools. The sampling procedure was the same. Fewer schools were sampled in the pilot because fewer forms were tested. However, the proportions of participating students by gender and ethnicity were very similar to that of the tryout. When a sampled school declined to participate in the pilot, a substitute school with similar characteristics was replaced. The number of students taking each form is listed in Table 19 below.

Table 19. Number of Students Participating in the Pilot by Form

Form	# of Students
12	1361
13	1340
14	1293
15	1178
16	1306
17	1320
18	1341
19	1209
Total	**10348**

Pilot Administration

Sampled schools were asked to test all eleventh grade students during a five-day administration window in April 1995. Classroom teachers were asked to administer the test. For security purposes and to minimize the exposure of test forms, makeup testing for students who were absent during the pilot was not recommended.

General Results

A summary of the descriptive statistics by form and by individual items is presented in Table 20 and 21 (Appendix C).

Table 20 provides descriptive statistics for both the complete sample that took a form and the two constituent subsamples taking the same form as it was administered within spiraled sets of two forms. Complete sample form means ranged between 31.18 (Form 13) and 34.99 (Form 16) out of 57-60 possible points. The mean p-values were between .54 and .60 on all of the test forms. This indicates that these items were moderately difficult for the 11th grade student sample. Considering each form as a whole, the mean item-test correlations were around .40 and the alpha coefficients were around .90 for all forms. Both of these statistics were very high, implying that the forms were very consistent internally.

The raw means and p-values are presented for all items in Table 21, Appendix C. In general, the distributions of p-values spread relatively evenly within a form. While this implies that the items were fairly distributed for this pilot sample, very few items had p-values below .20. The p-values

for some of the constructed-response items were, on average, lower than those of the multiple-choice items. This finding is not surprising in that it was the first time that constructed-response items were used on MEAP tests. In addition to individual item statistics, the gender and ethnic group descriptives and alpha coefficients for each of the five dimensions of science measured by the HSPT are presented in Table 21 (Appendix C).

Interrater Agreement

Scorers were hired and trained by CTB to score the pilot test constructed-response items using Michigan standards. The eight constructed-response items in each form were worth from one to three points each. On the pilot, scores for constructed-response items were obtained by averaging the ratings of two or three judges and rounding to the nearest integer. Only when the two readers' scores were not the same or adjacent - that is, more than one point apart on the same item - was the third reader introduced. Table 22 contains ranges for judges' agreement and consistency. Excluding those indices computed for judges who read very few papers (indicated in parentheses), agreement and consistency indices ranged between 88% and 100%.

Table 22. Interrater Agreement Ranges

FORM NUMBER	AGREEMENT RANGE (%)	CONSISTENCY RANGE (%)
12	88-100	88-100
13	94-100 [88(8)]*	96-100 [88(8)]
14	81-100	95-100
15	95-99 [88(8)]	96-99 [88(8)]
16	88-100	88-100
17	94-100 [88(8)]	94-100 [88(8)]
18	97-100	98-100
19	95-100 [88(8)]	96-100 [88(8)]

Agreement - percentage of times that a reader agreed, within one point, with the second reader.
Consistency - percentage of times that a reader agreed, within one point, with the second or third reader.
* One reader completed only eight readings for Form 13 with an agreement rate of 88%. The next lowest agreement rate for this form was 94%.

Additional reader interrater agreement statistics are presented in Tables 23-24 of Appendix C. The mean rate of exact agreement between the first two readers was at least 69% for all items (Table 23), with agreement ratios going down as the point values go up. There was no average non-adjacent reader agreement greater than 4%.

The frequency distributions of raw scores for the constructed-response items varied greatly within a form (Table 25, Appendix C). For example, on item 32 of Form 13, 690 students received zero points and only 194 students got the maximum number (2) of points possible. On item 50 of the same form, only 153 students received zero points, while 881 students scored the maximum number of points possible (2).

It should be noted that there were from 34 to 350 students choosing to leave the constructed-response items blank. Most constructed-response items had from 100 to 200 of the tested students not answering.

Group Descriptive Analyses

Descriptive statistics for four groups - whites, African-Americans, females, and males are presented in Table 26 for each of the eight science forms. Males had higher means than females on seven of eight science forms, while white means are higher than African-American means on all forms of the science test. The differences in group means were generally larger for the science and mathematics forms than for the reading forms.

African-American form means in Table 26 are based on less than 100 students for science Forms 14, 15, 16 and 19. The particularly low number, 43, of African-Americans students taking Form 15 is due to a school dropping out of the sample after agreeing to participate. The relatively small number of African-Americans in other forms may be attributed to the difficulty of getting high schools with large African-American enrollments to participate in the pilot.

Gender/Ethnicity DIF Statistics

Table 27 (Appendix C) contains DIF (differential item functioning) statistics, in the form of standardized mean differences (SMDs) for two group comparisons: males versus females and whites versus African-Americans. The SMDs for each comparison were partitioned into four groups in accordance with the procedure used for the tryout forms. Items that demonstrate large "practically significant" DIF against males or whites have SMDs greater than or equal to .20. Items that demonstrate "practically significant" DIF against females or African-Americans have SMDs smaller than or equal to -.20. A SMD between .10 and .19 (inclusive) or between -.10 and -.19 (inclusive) denotes items that have "practically significant" DIF against males and whites or against females and African-American students, respectively.

Given the magnitude of the SMDs for the items demonstrating large "practically significant" ($|SMD| \geq .20$) versus "practically significant" ($.10 \leq |SMD| \leq .19$) DIF, any item with a SMD in the former category can be considered to manifest twice the amount of ("practically significant") DIF against one of the four assessed groups than items with SMDs in the latter category. Hence a determination of the total amount of "practically significant" DIF that a form demonstrates against any one of these four groups can be obtained by multiplying the number of items manifesting large "practically significant" DIF by two and adding the number of items that demonstrate "practically significant" DIF. Note that several white versus African-American comparisons are based on relatively few (less than 100) African-Americans.

The eight science pilot forms were constructed, using the tryout DIF statistics, to ensure that the absolute difference in the amount of DIF (hereafter synonymous with "practically significant" DIF) of whites versus African-Americans **and** the absolute difference in the amount of DIF of males versus females was no greater than three. The purpose of constraining the absolute difference in DIF to no more than three for each of the two group comparisons was to ensure that DIF was relatively balanced across each of the two groups in each of the two comparisons.

The absolute difference in the amount of total DIF for the 16 comparisons (2 comparisons times 8 forms) can be seen in Table 27, within each pair of evaluated groups. The differences were frequently very small. For only one of the 16 comparisons does the absolute difference in DIF exceed three. This one comparison includes an absolute DIF of four against African-Americans for Form 14. The existence of one comparison that attained an absolute DIF difference greater than three in the pilot and not the tryout may most likely be attributed to the sampling variability of the tryout and pilot DIF statistics.

Summary

In summary, even though they were difficult, all the pilot forms showed high test reliability. Students had more difficulty answering constructed-response items than multiple-choice items. In fact, a fairly large proportion of students did not respond to the constructed-response items. The interrater agreement between the two scores was highest for the 1-point constructed-response items and lowest for the 3-point items.

Part 4. Student Survey and Teacher Survey

The Technical Advisory Committee (TAC) recommended that a study be done prior to the first administration of the Michigan High School Proficiency Test and again just prior to the time when the first graduating class would be impacted.

In early 1994, planning for an opportunity-to-learn study began. It was tentatively agreed that the final responsibility for the design must reside at the State Department level, that members of the Framework Committees should be involved in the design, that teachers in every district needed to be surveyed, that students should be sampled, and that the TAC should review the sampling plan and the draft survey instrument(s).

In March 1994, one TAC member, Department staff, and a member of the Science Framework Committee reached two major decisions:

(1) Surveys would be sent to every high school to the subject matter coordinators for the content areas tested on the HSPT. They would be asked to form committees of teachers from their high schools as well as their feeder schools to fill out the survey.

(2) A sample set of students would be part of the study.

In subsequent meetings with the Science Framework Committee, discussions were held regarding the content and the format of the surveys. It was agreed that the general form of the surveys was to be the same across content areas, but that form should not take precedence over substance and if there were good reasons for having different formats, it would be allowed. Content area experts were to be responsible for the actual wording of the surveys.

The study was originally intended to address three purposes: (1) to help make adjustments to the tests if necessary, (2) to aid in standard setting and (3) to provide schools with information that could be used for professional development.

On September 2, 1994, an overview of the proposed design was presented to the TAC. The TAC members suggested that the names of the surveys be changed from "opportunity-to-learn" surveys to the "Teacher Survey" and the "Student Survey." Revisions were suggested and made for the Student Survey. The Teacher Survey was discussed at length, reviewed and revised. Both the student and teacher surveys were piloted at several sites before being sent out.

Science Student Survey Results

The Science Student Survey (Appendix D) was given to the students who participated in the science tryout. The students completed the survey prior to taking the item tryout "test" so that student perceptions pertaining to performance would not influence survey responses.

The science survey contained 29 statements. The common stem for the first 16 questions was as follows: "By the end of tenth grade, how often did your school experience include:..." For questions 17 - 29, students were asked to "estimate how often you studied each topic by the end of tenth grade." Students were to respond on a four-point scale from "never" to "a lot." Note that "never" was translated to a value of "zero" (0), "very little" to "1," "some" to "2," and "a lot" to "3."

Table 29 below presents the summary data for the student survey. The mean score for the 29 science survey questions was 1.77 (2 = some). The lowest mean for a survey question (#13) was 0.99, which was the only question with a mean below 1.00. Eight questions (28%) had a majority

of the students respond "less than some" (2). Nine questions (31%) had a mean less than 1.5. The science results are probably the least positive of the content areas.

Because the surveys were given to the same students who participated in the tryout, it was possible to correlate the mean scores for the students on the survey with their scores on the tryout tests. The correlations are positive, but not particularly high (.1706). Thus, the students' perceptions of whether they were taught something did not seem very highly related to how they actually scored on the tryout.

Table 28: Student Survey Results Summary
Content: Science

Total # of questions	27
Overall mean	1.77
Lowest mean	0.99
# & % of questions that the majority marked less than "some" (2.0)	8 (28%)
# & % of questions with a mean less than 1.5	9 (31%)
Correlation statistic of survey mean and tryout score	.17

Conclusions From the Student Survey

In drawing conclusions from the student survey results, one must keep in mind that there was no good way of determining how honestly students responded to the questions or even the extent to which they understood the questions. Given those cautions, it was concluded that school experiences in general included the types of activities useful in assisting students to learn the content to be tested on the proficiency test. Generally, students' responses indicated that the activities were experienced more than "very little."

Science Teacher Survey

The Teacher Survey was sent to science supervisors at all high schools in the state (N=758), May of 1995. These supervisors were each to form a team of teachers to work with them in completing the Teacher Survey and an Instructional/Curriculum Support Materials Form, which they did not need to return.

The science teacher survey is composed of 91 statements organized by scientific dimensions, objectives and outcomes within dimensions. The dimensions are as follows: (a) using life science, (b) using earth science, (c) using physical science, (d) constructing science knowledge, and (e) reflecting science knowledge. For each statement, the respondents completed two columns. In the first column, they circled all grades receiving instruction, and in the second column they circled the

one grade at which sufficient classroom instruction had occurred to expect understanding/ proficiency.

Summary Of the Teacher Survey Results

In summarizing the science teacher survey results, it must be remembered that the data analyzed were based on a low return rate of 244 responses out of 758 surveys sent to schools and may not be representative. Nevertheless, some tentative findings emerge from the teacher survey results that are summarized in Table 29:

- only one statement had more than 25% of the schools circle Not Taught (NT);
- only one statement had 50% or more of the schools circle Not Sufficient Instruction (NSI);
- seventy-three statements had fewer than 25% of the schools circle "NSI"; and
- seventeen statements had "NSI" circled by fewer than 10% of the schools.

Table 29. Teacher Survey Results Summary
Content: Science

# and % of statements where NT circled by 25% or more	1 (1%)
# and % of statements where NSI circled by 50% or more	1 (1%)
# and % of statements where NSI circled by 25% or more	18 (20%)
# and % of statements where NSI circled by less than 10%	17 (19%)

Overall Summary And Follow-Up[3]

Both the student and teacher survey results suggested that many of the objectives were already being taught in the majority of the schools and that they were sufficiently taught for students to have proficiency in them. However, in science, there were a number of objectives that were not judged to have been taught with sufficient thoroughness.

The results of both the teacher and student surveys were presented to the standard setting committees at the time they made recommendations regarding scores. Prior to that time, the department devoted considerable time determining just how the data should be presented and what the committees should be told about the relevance of the data for standard setting. It must be stressed that these data were gathered in the 1994-95 school year, and that information about the content of the proficiency tests continued to be widely disseminated before the test was given in the

[3] In July, 1996, the State Board of Education approved the standards as set by the standard setting committees, without changes. Information about the student and teacher surveys is adapted from a 1996 paper presented by Mehrens, Smolen and Yan at the Michigan School Testing Conference, Ann Arbor, MI.

spring of 1996. It is reasonable to believe that instruction in the schools has become more aligned to the objectives tested as time has passed.

The results of these surveys were disseminated to curriculum coordinators in the schools who were encouraged to use them in planning curricular/instructional changes prior to the first administration of the HSPT. It should have been clearly understood by local schools that it is in the best interests of their students to teach them material from a content domain that is sampled on a test for which passing is a requirement for a state-endorsed certificate.

REFERENCES

American Association for the Advancement of Science (1989). Science for All Americans. New York, NY: Oxford University Press, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Burket, G. R. (1991; 1995). PARDUX. Monterey, CA: CTB Macmillan/McGraw-Hill.

Linacre, J. M. & Wright B.D. (1993). A users guide to BIGSTEPS: Rasch model computer program. Chicago, IL: MESA.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association, 58, 690-700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of a disease. Journal of the National Cancer Institute, 22, 719-748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Mehrens, W. A., Smolen, D. L., & Yan, J. W. (1996). Michigan High School Proficiency Test. Summary of Student and Preliminary Teacher Survey Results. Paper presented at the 1996 Michigan School Testing Conference. Ann Arbor, MI.

Michigan State Board of Education (1980). The Common Goals of Michigan Education. Lansing, MI.

Michigan State Board of Education (October, 1991). Model Core Curriculum Outcomes. Lansing, MI.

Michigan State Board of Education (1991). Michigan Essential Goals and Objectives for Science Education. Lansing, MI.

Michigan State Board of Education (1994). The Assessment Framework for the Michigan High School Proficiency Test in Science. Lansing, MI.

Muraki, E. (1992). A generalization partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Stocking, M., & Lord, F. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 5, 245-262.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30, 233-251.

Appendix A

Expert Panel*

Mr. Thomas Fisher
Administrator of Student Assessment Services Section
Florida Department of Education

Ms. Sharon Johnson-Lewis
Director of Planning, Research and Evaluation
Detroit Public Schools

Ms. Marjorie Mastie
Supervisor for Assessment Services
Washtenaw Intermediate School District

Dr. William Mehrens, Expert Panel Chair
Professor of Educational Measurement
Michigan State University

Dr. Jason Millman
Professor of Educational Measurement
Cornell University

Dr. S.E. Phillips
Associate Professor of Education
Michigan State University

Dr. Edward Roeber
Director of Student Assessment Programs
Council of Chief State School Officers

Dr. Roger Trent
Director, Division of Educational Services
Ohio Department of Education

* Job titles at time panel convened.

Technical Advisory Committee (TAC)*

Dr. Gail Baxter
Assistant Professor of Education
University of Michigan

Dr. Roger Trent
Director of Assessment and Evaluation
Ohio Department Of Education

Ms. Sharon Johnson-Lewis
Assistant Superintendent
Research, Development and Coordination
Detroit Public Schools

Dr. William Mehrens
Professor of Educational Measurement
Michigan State University

Dr. Edward Roeber
Director, Student Assessment Programs
Council of Chief State School Officers

Dr. Joseph Ryan
Research Consultant Center
Arizona State University West

* Job title at time of HSPT development

Exercise Development Team (EDT) - Science*

Dr. Richard Fidler
Science Teacher, Department Chair
East Junior High School
Traverse City

Dr. Tim Falls
Principal, Meadows Elementary School
Novi Community Schools

Mrs. Annis Hapkiewicz
Chemistry, Science Coordinator
Okemos High School
Okemos School District

Mr. David Kraepel
Science Teacher
Monroe Junior High School Science
Detroit Public Schools

Ms. Susan Krussel
Science Teacher
Rochester High School
Rochester School District

Mr. Ted Lau
Physical Science Teacher
Northwestern Senior High School
Jonesville School District

Dr. Raymond Leising
Biology, Senior High Science Teacher
Jonesville High School
Jonesville School District

Mr. David Mastie
Earth Science Teacher
Pioneer High School
Ann Arbor School District

Mr. Henry Thoenes
Science Teacher
Boulan Park Middle School
Troy School District

* Job title at time of HSPT development

Content Advisory Committee (CAC) - Science*

Ms. Barbara Berthelsen
Science Coordinator
Big Beaver Center
Troy School District

Ms. Sally DeRoo
Special Education, K-9 Science
Plymouth Schools

Dr. Don Collins
Science Coordinator
Flint School District

Dr. Tim Falls
Principal, Meadows Elementary School
Novi Community Schools

Mr. Richard Gaubatz
Principal, Whitmore Lake High School
Whitmore Lake Public Schools

Dr. Dave Housel
Director, Oakland School Science Center
Oakland Intermediate School District

Mr. Ron Kaminskis
Science Teacher
Scarlett Middle School
Ann Arbor School District

Dr. Richard LeFebre
Professor, Geology Department
Grand Valley State University

Mr. Gary Cieniuch
Science Coordinator
Livonia Public Schools

Mrs. Sarah Lindsey
Coordinator of Science
Midland Public Schools

Dr. Mike Marlowe
Math/Science Coordinator
Jackson Intermediate School District

Dr. Howard Stein
Professor, Biology Department
Grand Valley State University

Mr. Henry Cole
Coffey Middle School
Detroit School District

* Job title at time of HSPT development

Bias Review Committee (BRC)*

Ms. Ellen Carter-Cooper
Educational Consultant/
School Development Unit
Michigan Department of Education

Dr. Rossi Ray-Taylor
Director of State and Federal Programs
Lansing School District

Ms. Marian Phillips
(replaced Dr. Ray-Taylor)
Supervisor, Research and Evaluations
Lansing School District

Mr. Aden D. Ramirez
Director, Bilingual/Migrant Program
West Ottawa Public Schools

Ms. Stephanie Rockette
Mathematics Resource Teacher
Vincent Place/Teacher Resource
Benton Harbor Area Schools

Dr. Elana Izraeli, District Coordinator
Testing & ESL Programs
West Bloomfield School District

Mr. H. William Leavell, Jr.
Research Specialist
Michigan Jobs Commission
Michigan Rehabilitation Services

Dr. Pauline Coleman
English Language Arts Coordinator
Ann Arbor Public Schools

Mr. Robert Brown
Huron High School
Ann Arbor Public Schools

Mr. Jesus M. Solis
Educational Consultant
Michigan Department of Education

Mr. William Gay
Teacher/Huron High School
Ann Arbor Public Schools

* Job title at time of HSPT development

Expert Panel Recommendations

1. The State Board should not specify subject areas other than Communications Skills, Mathematics, and Science for the initial assessment.

2. Communication skills assessed during the first assessment cycle should be limited to reading and writing.

3. The State Board and the Michigan Department of Education need to determine which subsets of the model core curriculum should be included in the assessments. This needs to be done very shortly. The decision should be based on recognition of the importance of students' opportunity to learn the content and some knowledge regarding what is likely to be in the school curricula by the date of the first test. The decision should not be that the total core curriculum is the appropriate domain from which to build the tests.

4. Once a determination is made regarding the testable portion of the core curriculum, there should be an administrative rule or statute that specifies this portion of the core is exempted from the permissive language in P.A. 25 and must be taught by the local districts to all students.

5. Once the testable portion of the core is determined, there should be wide publicity of this to the local districts. Consideration should be given to how this information can be disseminated with enough detail to let students and educators know the knowledge and skills to be tested but without so much detail that the students can answer the questions without understanding the curricular elements from which the items are only a sample.

6. Gather evidence from both teachers and students regarding the opportunity to learn the content domain the tests sample prior to the first administration.

7. Provide instructional support and training to local teachers if there is a need.

8. The State Board should not make any changes in the core curriculum or selected testable core prior to 1997.

9. When (or if) any changes are made in the core curriculum, there must be a phase-in period, and the tasks described in recommendations 3 through 7 would need to be repeated.

10. Name the assessment the "Michigan High School Graduation Tests."[4]

11. The Department of Education should caution its employees and the State Board against making any unsubstantiated statements about what the tests measure or what inferences can be made from the test scores. There should be an official statement about the tests and the inferences that can be drawn from the scores.

12. Demand that the test developer design sufficient safeguards to ensure that the test adequately samples the defined content.

13. Be careful not to make any official statements that would suggest the test has criterion-related validity if supportive data have not been gathered.

---

[4] Because there will be different tests for different content areas, we suggest the plural "tests." However, for ease in subsequent writing we will, at times, refer to the total assessment as a test. When we do so, it should be understood that the reference includes all the tests.

14. Contract for enough items initially so that after losses through pilot and field testing there will be enough to build forms through the 95-96 administration year.

15. Reissue a contract in sufficient time to have items developed and tried out (possibly embedded in a live form) prior to their being needed for the 96-97 year.

16. Schedule a large scale field tryout for tenth graders by the spring of 1994.

17. Appoint and train a standard-setting committee.

18. Use a technical advisory committee to help develop a specific standard-setting procedure.

19. The State Board of Education should establish a passing score through administrative rule based upon a recommendation by the superintendent of public instruction with the advice of appropriate committees.

20. Consider setting incremental cut scores for different graduating classes at the time the State Board of Education makes its initial decision.

21. The item sensitivity reviews should be completed by a committee that is selected and trained specifically for this task. Most members should represent Michigan's predominant minority groups. However, it would be wise to have at least one member of the committee be a minority group member from out-of-state who is a recognized expert in the area.

22. Statistical item bias studies should be conducted. Items which show up as statistically biased should be reviewed (but not necessarily discharged) by an item bias committee (conceivably, but not necessarily the committee used for the item sensitivity review) and a content review committee.

23. Obtain the following reliability estimates: internal consistency, inter-rater reliability, generalizability across writing samples, and the reliability or standard error at the cut score.

24. Scores should be reported as "Pass" or "Fail." Those individuals who fail should be given some information regarding how close they were to passing, and they should be given some diagnostic information that would facilitate remediation efforts. There are important technical details (e.g., reliability of difference scores) regarding various methods of reporting diagnostic information and specific plans should be formulated by a technical advisory committee prior to approval of the final test specifications.

25. We would encourage use of a common scale across subject matter areas. This takes some advance planning to avoid adopting a scale that is appropriate for one test, but unworkable for another.

26. Develop detailed rules (procedures) for designating forms for make-up examinations and out of school (i.e., Adult Ed.) populations. Determine whether you should ever reuse a form. Determine how many times you will administer the test each year. Determine equating procedures (e.g., number of anchor items to be used). Based on these considerations, initially develop enough alternate forms to last through at least the 1995-96 school year. Start developing more forms/items prior to that so a sufficient supply is continuously available.

27. Use a technical advisory committee to help develop specific equating procedures.

28. Consider carefully policies regarding all test administration conditions. For example, the decision of whether or not to use calculators in the mathematics test must be made by the department, not by local school personnel. Train local school personnel adequately to administer the tests. Consider random auditing of the administration process to ensure uniformity throughout the state.

29. Be cautious about any "predictive" interpretation of the scores of any single individual from testing in earlier grades. Such tests should be thought of as providing only an early awareness.

30. The department should prepare and have the board adopt written procedures regarding make-up examination provisions.

31. The department should prepare and have the board adopt specific written rules regarding the number of retakes that should be allowed, and how many attempts a student should be given prior to the time he/she is scheduled to graduate.

32. Develop a detailed proposal that addresses questions regarding remediation efforts and the respective responsibilities of the state, the district and the student for remediation efforts.

33. Enact an administrative rule regarding testing issues related to special education students and students with limited English proficiency.

34. Individuals in adult education programs who wish to receive high school diplomas after the end of the 1996-97 school year should be required to pass the High School Graduation Test.

35. Obtain the services of the Attorney General's Office early on in the process and continuously as new policies are developed and implemented.

36. The State Superintendent of Public Instruction and the State Board of Education should work with the legislature to adopt statutory authority for the high school graduation testing program.

37. Carefully investigate liability issues with assistance from the Attorney General's Office. Attempt to obtain necessary statutes with respect to liability. Inform all committees and all staff regarding their potential liability.

38. Schools should be notified immediately regarding this graduation requirement and the information disseminated to all teachers. Students and their parents should be notified no later then the spring of 1993.

39. The department should prepare, and the board should adopt, detailed policies regarding what should be documented and how long the documentation should be kept on file. We generally suggest that all documentation be kept for a period of at least five years following the school year in which the test was administered. We suggest keeping "forever" the initial development documentation and records about when, why, and how procedures are adopted and/or changed.

40. In consultation with the Attorney General's Office, and based in part upon discussions with representatives of state education associations (e.g., teachers' unions and administrators' associations), the department should prepare, and the State Board of Education should adopt, rules regarding what constitutes inappropriate behavior on the part of educators or students with respect to test-taking behavior, security issues, and so forth; and what

penalties will be imposed for violation of these rules. These rules and the penalties should be disseminated to educators and students prior to the initial administration of the graduation test.

41. The department needs to develop a complete list of rules/regulations that need to be adopted and decide whether these can simply be adopted by the board or whether they need legislative approval.

42. Detailed security arrangements need to be developed.

43. Detailed policies regarding security valuations need to be established. Staff should investigate current laws regarding freedom of information exclusions, and if they are insufficient, request new legislation to exempt secure test materials from the freedom of information regulations.

44. The department needs to determine what additional equipment/facilities are needed for storage of secure materials, shredding out-of-date secure materials, etc.

45. An annual test administration plan should be developed and disseminated to all school districts.

46. The tests should first be administered to 10th graders in the spring of 1995 and they should be administered at least twice each in the junior and senior years.

47. The department should conduct a careful study to assess additional staffing needs in assessment and instructional programs.

48. The position of supervisor of state assessment should be filled as quickly as possible.

49. The following advisory committees should be appointed: 1) a Michigan Department of Education Steering Committee, 2) a Testing Policy Advisory Committee, 3) a Bias Review Panel, 4) a Technical Advisory Committee, 5) a Content Review Committee in each content area of the test, 6) an overall content review committee, and 7) a Standard Setting Committee.

50. Use at most two contractors: one for test development and formal field tryouts; and another for test administration, scoring, and reporting.

51. Obtain more detailed information from other states with similar programs regarding fiscal needs. Make recommendations to the legislature that are sufficient to cover department needs, and make clear to them that the task simply cannot be done without adequate support.

BIAS REVIEW COMMITTEE COMMENT SHEET

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM
MICHIGAN DEPARTMENT OF EDUCATION

TEST ITEMS BEING REVIEWED (Content Area and Grade) _____

DATE_____ MDE Representative_____

The below items were judged to be problematical by the Bias Review Committee.

Form #	Item #	Bias Issue	Comments

BIAS REVIEW COMMITTEE

MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM
MICHIGAN DEPARTMENT OF EDUCATION

TEST ITEMS BEING REVIEWED (Content Area and Grade) _____

DATE_____ MDE Representative_____

The below items were reviewed by the Bias Review Committee who were asked to review items with sensitivity to gender, racial or ethnic groups, religious groups, socioeconomic groups, people with disabilities, and regional concerns in mind. Checked items were not judged to be biased in the above categories. Items with an asterisk (*) were judged to be biased and therefore have further comment and explanation on the attached Bias Review Comment sheet.

Form #	Item #	Stereotyping	Underlying Assumptions	Controversial Material	Contextual Concerns	Elitism or Ethnocentrism	Other

Michigan School Stratum Classification

The Michigan schools are classified into seven strata relative to populations where the schools reside.

1. Large City
   Central city of a Metropolitan Statistical Area (MSA) with a population greater than or equal to 400,000 or a population density greater than or equal to 6,000 people per square mile.

2. Mid-size City
   Central City of an MSA with a population less than 400,000 and a population density less than 6,000 people per square mile.

3. Urban Fringe of Large City
   Place within an MSA of a Large Central City and defined as urban by the Census Bureau.

4. Urban Fringe of Mid-size City
   Place within an MSA of a Mid-size Central City and defined as urban by the Census Bureau.

5. Large Town
   Town not within an MSA and with a population greater than or equal to 25,000 people.

6. Small Town
   Town not within an MSA and with a population less than 25,000 and greater than or equal to 2,500 people.

7. Rural
   A place with fewer than 2,500 people and coded rural by the Census Bureau.

Criteria for Writing and Editing Multiple-Choice Items

- ❒ The item is free of gender, ethnic, racial or other bias.
- ❒ The content of the item is grade-appropriate.
- ❒ The reading level of the item stem and answer choices is suitable for the student being tested.
- ❒ All factual information has been checked and documented against reliable, up-to-date sources.
- ❒ A student possessing the skill being tested can clearly select one and only one correct response.
- ❒ All extraneous material has been edited from the stem.
- ❒ All item distracters are plausible to someone who has not mastered the skill being measured.
- ❒ Answer choices are free of repetitious words or expressions that can be included in the stem.
- ❒ All answer choices are consistent with the stem both conceptually and grammatically as well as consistent with each other.
- ❒ All answer choices are mutually exclusive.
- ❒ All answer choices in the item are approximately equal in length (i.e., no one choice is much longer or shorter than another).
- ❒ No outliers - answer choices that are obviously different from the others.
- ❒ The correct response for the item has been indicated.
- ❒ Art has been conceptualized and sketched for the item, if applicable.
- ❒ The passage/stimulus associated with the item has been provided.

Checklist for Item Development

- ❒ The item matches content and format specifications.
- ❒ The item deals with material that is important in testing the appropriate strand.
- ❒ The item is free of gender, ethnic, racial, or other bias.
- ❒ The content of the item is grade-appropriate.
- ❒ The thinking skills demanded of the student are grade-appropriate.
- ❒ The reading level of the item strand and answer choices are suitable for the student being tested.
- ❒ All factual information has been checked and documented against reliable, up-to-date sources.
- ❒ The student can answer the question or complete the statement without looking at the answer choices.
- ❒ A student possessing the skill being tested can clearly select one and only one correct response.
- ❒ All item distracters are plausible to someone who has not mastered the skill being measured.
- ❒ The item stem presents only one question or statement.
- ❒ The item stem does not present clues to the correct response of the item.
- ❒ The item (stem and/or answer choices) does not present clues to the correct response to any other item that is in the same set of choices.
- ❒ All extraneous material has been edited from the stem.
- ❒ Answer choices are free of repetitious words or expressions that can be included in the stem.
- ❒ All answer choices are consistent with the stem both conceptually and grammatically as well as consistent with each other.
- ❒ All answer choices in the item are approximately equal in length (i.e. no one choice is much longer or shorter than another; in math, from low to high or vice-versa).
- ❒ All answer choices are mutually exclusive.
- ❒ No outliers (responses that are obviously different from the others):
  - ❒ Responses all similar in meaning.
  - ❒ Responses either all similar in length or two are long and two are short.
- ❒ Answer choices should not all begin with the same word - if this happens, include the word or words in the stem.
- ❒ Items phrased clearly and simply (check words that you suspect are too difficult a reading level against some word list).
- ❒ Check for similarity of items, repeated items, or items that give clues to other items.
- ❒ Check whether any material is copyrighted and, if so, indicate source so permission can be obtained.

- ❒ Reasonable representation of economic classes, races, ages, sexes, and handicapped in text and art:
  - ❒ Variety of above graphics.
  - ❒ Non-stereotypic representation.
  - ❒ Watch middle- and upper-economic level bias.
- ❒ Check to see that opinions are not masquerading as facts.
- ❒ Junk food?
- ❒ Is the material too dated for audience?
- ❒ The negative form of the stem has been used only if absolutely necessary.
- ❒ Key words (e.g., best, first, not, etc.) are formatted according to specifications (underlined, capitalized, italicized, left alone).
- ❒ The correct response for the item has been indicated.
- ❒ Art has been conceptualized for the item, if applicable.
- ❒ Position and type of art is indicated.
- ❒ Each piece of art is described in words and/or pictures.
- ❒ Descriptions of each piece of art are specific and unambiguous.
- ❒ Rules are clear, straight, of desired width and length. Sides drawn proportionally.
- ❒ Art has been checked against the corresponding item. Art or item has been revised, if necessary.
- ❒ Figures and tables are accurate, factual, and documented if appropriate.
- ❒ Males and females are represented equally in the art.
- ❒ Ethnic groups are represented equitably and non-stereotypically in the art.
- ❒ The passage/stimulus/graphic associated with the item has been indicated.

NOTE: Use your project checklist in addition to this checklist.

Sign Off

_____ _____

Name Date

Checklist for Scoring Rubrics/Scoring Guide

- ❒ Type of scoring for each scorable unit has been identified.
- ❒ A scoring rubric has been identified for each scorable unit prior to or simultaneously with item development.
- ❒ The performance criterion (outcome/strand to be assessed) has been identified for each scorable unit.
- ❒ All foreseeable correct responses have been identified.
- ❒ A scale (no. of points) has been identified for each scorable unit.
- ❒ Score points have been defined for each scorable unit (e.g., 4 = outstanding).
- ❒ Score points are clearly distinguishable from one another.
- ❒ The rubric allows full credit for answers dependent on earlier responses, even if the earlier response is incorrect.
- ❒ When more than one student behavior is required by an activity, the rubric clearly distinguishes among the behaviors and indicates how each is to be scored.
- ❒ The rubric focuses on performance (i.e., what the student did) and not on the performer (i.e., what the student understands).
- ❒ The language of the rubric is clear, consistent, and unambiguous.
- ❒ Any changes to scoring rubrics have been checked against the corresponding item.
- ❒ Scoring rubrics have been revised if any revisions occurred in the corresponding item.

Sign Off

_____ _____

Name Date

Appendix B

Table 9. Michigan HSPT in Science Tryout
Raw Score Statistics by Form

Form	Grp	# of Scored Items	N	Raw Score			α	P-Value[1]		IT[2]		Collapsed Levels		
				Mean	%MS[3]	SD		90th	10th	90th	10th	Item #	From	To
20	1	54	514	36.8	46	15.7	.90	.78	.31	.50	.16	--	--	--
	5	54	521	41.3	52	14.3	.88					--	--	--
21	1	53	507	33.3	42	14.3	.88	.78	.24	.51	.06	--	--	--
	5	53	541	37.1	46	14.0	.88					--	--	--
22	1	54	507	32.8	40	14.9	.89	.68	.29	.52	.07	15	7	6
	5	54	538	32.6	40	15.0	.89					15	7	6
23	2	54	519	35.3	44	14.4	.90	.81	.28	.47	.16	32	7	6
	6	54	544	34.1	43	14.3	.90					32	7	6
24	2	54	526	38.6	49	15.6	.91	.78	.32	.52	.19	--	--	--
	3	54	477	39.2	50	14.0	.89					--	--	--
25	3	53	471	33.8	42	11.8	.86	.72	.27	.42	.12	--	--	--
	6	53	545	33.5	42	13.3	.89					--	--	--
26	3	54	469	37.4	46	14.2	.87	.80	.22	.50	.16	--	--	--
	4	54	404	36.8	45	13.8	.87					--	--	--
27	4	53	395	37.1	46	15.1	.90	.76	.28	.50	.09	--	--	--
	6	53	549	35.9	45	15.3	.91					--	--	--
28	4	54	405	35.9	44	15.6	.89	.67	.33	.56	.14	--	--	--
	6	54	549	36.1	45	16.8	.91					--	--	--
29	4	54	403	36.3	47	13.9	.89	.77	.26	.49	.12	--	--	--
	5	54	531	37.7	48	14.1	.89					--	--	--

1. P-values for 90th and 10th percentile when items are sorted in order of p-values.
2. Items/test correlations for 90th and 10th percentile items.
3. Mean divided by maximum score (percentage of maximum score).

Table 10. HSPT in Science Tryout
Summary of Fit Results - 1PL/PPC

Grp	Form	N	# of Scored Items	# of Misfit Items				Two largest Zs		Unest. Items	
				Z≥10	10>Z≥5	5>Z≥3	3>Z≥2			Number	Item #
1	20	514	54	3	7	8	10	28.4	10.9	0	
5		521	54	2	11	7	6	31.5	11.9	0	
1	21	507	53	3	7	8	6	18.4	13.3	0	
5		541	53	4	8	9	8	15.1	14.4	0	
1	22	507	54	3	6	12	7	36.6	18.9	0	
2		538	54	6	1	10	10	27.3	21.7	0	
2	23	519	54	4	5	8	6	24.4	17.1	0	
6		544	54	3	7	8	9	25.5	19.8	0	
2	24	526	54	3	6	6	11	15.0	13.9	0	
3		477	54	2	7	9	6	14.3	12.0	0	
3	25	471	53	2	6	9	6	12.0	10.3	0	
6		545	53	3	5	7	10	61.1	19.1	0	
3	26	469	54	2	5	7	6	14.7	14.1	0	
4		404	54	3	4	9	3	16.0	14.8	0	
4	27	395	53	3	4	8	7	30.0	19.8	0	
6		549	53	6	5	12	7	28.4	22.6	0	
4	29	403	54	2	7	7	8	23.5	10.9	0	
5		531	54	2	9	9	2	22.9	12.3	0	

Table 11. HSPT in Science Tryout
Summary of Fit Results - 3PL/2PPC

Grp	Form	N	# of Scored Items	# of Misfit Items				Two largest Zs		Unest. Items	
				Z≥10	10>Z≥5	5>Z≥3	3>Z≥2			Number	Item #
1	20	514	54	0	0	0	4	2.7	2.1	1	7
5		521	54	0	0	1	3	4.6	2.3	4	3*, 7, 20*, 42
1	21	501	53	2	0	3	8	25.9	18.8	4	1, 29, 40, 42
5		540	53	0	1	2	3	6.4	4.5	4	1, 29, 40, 42
1	22	502	54	0	0	1	1	3.4	2.9	4	6, 18, 41, 52
2		537	54	0	0	1	6	3.6	2.7	2	18, 46
2	23	517	54	0	0	1	5	3.8	2.9	2	40, 43
6		547	54	1	0	3	3	56.9	3.9	1	43
2	24	525	54	0	0	1	5	4.9	3.0	0	
3		477	54	0	0	2	6	3.7	3.4	1	28*
3	25	471	53	0	0	1	4	4.9	3.5	1	5
6		534	53	1	0	1	5	21.0	3.8	2	5, 21
3	26	469	54	0	1	1	6	6.1	4.1	3	10, 26*, 51
4		404	54	0	1	1	4	6.3	3.5	4	10, 26*, 28*, 51
4	27	395	53	0	0	2	1	4.0	3.9	3	21*, 24, 50
6		542	53	0	1	3	5	5.6	4.7	3	24, 39, 50
4	28	405	54	0	0	0	1	2.2	---	2	22, 28
6		543	54	0	0	2	2	3.6	3.2	1	22
4	29	403	54	0	0	1	1	3.3	3.0	3	6*, 25, 42
5		530	54	0	1	0	2	5.6	2.6	3	6*, 25, 42

* Item/test correlation ≥ .08.

Table 12. HSPT in Science Tryout
Items Flagged for Deletion Under the Fit Criteria

	1PL/PPC		3PL/2PPC		
Form	# Misfit Items[1]	Item Number[2]	# Misfit Items[1]	Item Number	NC[3]
20	11	3, 7[$], 9, 20, 26, 29, 38, 39, 42[$], 43, 53	0		4
21	11	1[$], 6, 10, 21[$], 26, 29[$], 40, 42[$], 45, 46, 52	0		4
22	9	6[$], 18[$], 22[$], 26, 40[$], 45, 46[$], 48, 52[$]	0		4
23	9	1, 8, 21[$], **34**, 37, 40[$], 41, 43[$], 52	0		2
24	9	9, 11, 25, 28, 30, **33**, 36, 42, 46	0		1
25	7	3, 4, 5[$], 21[$], 26, 39, 40	0		2
26	6	10[$], 28, 30, **32**, 41, 51[$]	0		4
27	9	4, 12[$], 21, 22 24[$], 25, **33**, 39[$], 50[$]	1	22	3
28	6	3, 22[$], 38[$], **32**, 39, 50	0		2
29	10	14, **15**, 25[$], 26, 35, 41, 42[$], 43, 44, 48[$]	0		3

1. Note that each item has two Zs, one from one sample and the other from a second sample. A "misfit" item is defined as follows:
   (1) both Zs≥ 4.0, or
   (2) (one Z≥4.0), and (4.0> the other Z≥3.0), and a plot of expected and observed curves fails to demonstrate reasonable fit.

   Of the 87 items that were not fitted by the one-parameter model, 18 items fell into the latter category, (2). The single item that was not fitted by the 3PL/2PPC model also fell into the latter category.

2. Bold numbers indicate constructed-response items.

3. Maximum number of non-convergent items in a given form taken by two samples.

$. Item-test correlation < .08 signifying low discrimination.

Table 13. HSPT in Science Tryout
Means and Standard Deviations of
Discrimination: 3PL/2PPC

		All Items			Multiple-Choice Only			Constructed-Response Only		
Form	Group	# of Items	Mean	SD	# of Items	Mean	SD	# of Items	Mean	SD
20	1	50	1.53	0.72	42	1.69	0.67	8	0.68	0.14
20	5	50	1.36	0.69	42	1.52	0.63	8	0.53	0.12
21	1	49	1.26	0.64	41	1.37	0.64	8	0.71	0.15
21	5	49	1.33	0.73	41	1.48	0.70	8	0.56	0.08
22	1	49	1.34	0.66	41	1.47	0.64	8	0.64	0.12
22	2	49	1.41	0.75	41	1.56	0.74	8	0.67	0.13
23	2	52	1.30	0.57	44	1.41	0.55	8	0.70	0.18
23	6	52	1.31	0.61	44	1.42	0.60	8	0.71	0.16
24	2	53	1.47	0.60	45	1.60	0.56	8	0.78	0.23
24	3	53	1.32	0.57	45	1.44	0.54	8	0.69	0.26
25	3	51	1.33	0.82	43	1.45	0.83	8	0.67	0.20
25	6	51	1.28	0.58	43	1.38	0.57	8	0.71	0.13
26	3	50	1.25	0.62	42	1.37	0.61	8	0.64	0.19
26	4	50	1.30	0.71	42	1.43	0.70	8	0.63	0.20
27	4	49	1.33	0.58	41	1.45	0.56	8	0.75	0.23
27	6	49	1.34	0.55	41	1.45	0.53	8	0.79	0.24
28	4	52	1.41	0.63	44	1.53	0.61	8	0.77	0.19
28	6	52	1.28	0.51	44	1.37	0.50	8	0.76	0.13
29	4	51	1.39	0.68	43	1.52	0.66	8	0.72	0.21
29	5	51	1.30	0.64	43	1.40	0.63	8	0.73	0.21
Total All Forms		1013	1.34	0.65	853	1.47	0.63	160	0.69	0.19

Michigan High School Proficiency Test
Science Tryouts
Teacher Comment Sheet

As part of the Michigan HSPT Science tryout, the Michigan Department of Education is asking you to complete the following comment sheet.

Directions: Please answer each of the following to the **BEST** of your ability. Each item can be answered by the person administering the HSPT Science tryout. None of the items are specific to any particular form. IF YOU NEED MORE SPACE TO RESPOND, PLEASE USE THE BACK OF THESE SHEETS OR ATTACH YOUR OWN.

1. Was the Administration Manual clear, easy to use, and complete? _____Yes _____No
   If "no," what changes would you suggest?

2. Did you have a sufficient number of test materials? _____Yes _____No

3. Approximately what percentage of your students finished during the two hour block of time?
   _____ %

4. Did the students have any difficulty with the directions for the test? _____Yes _____No
   If "yes", please be specific.

5. Were there any charts, graphs or pictures that were not clear to the students?
   _____Yes _____No If "yes," please be specific.

6. Did the students raise any particular concerns about the constructed-response items?
   _____Yes _____No If "yes," please be specific.

7. Did the reading level of the test seem appropriate for grade 11 students?
   _____Yes _____No If "yes," please be specific.

8. Were there particular questions in any part of the test on which a large number of students had difficulty?
   _____Yes _____No If "yes," please be specific.

9. Were there other aspects of this test which gave the students, or you as test administrator, difficulty?

_____Yes _____No If "yes," please be specific.

10. In this section, provide your ideas, critique, etc., on this tryout. Please include student reactions to exercises as well as your overview of the entire test.

THANK YOU FOR YOUR TIME AND EFFORT IN RESPONDING TO THESE QUESTIONS.

Appendix C

Table 20. HSPT in Science Pilot
Descriptive Statistics by Form

Form	Set of Pilot Group	# of Scored Items	# Points	Mean	s.d.	N	α	P-Value[1] Mean	P-Value[1] s.d.	Item-Test Correlation Mean	Item-Test Correlation s.d.
12	-	50	59	34.48	10.74	1361	.90	.58	.18	.41	.13
	1	-		33.65	10.89	656	-	-	-	-	-
	4	-		35.26	10.53	705	-	-	-	-	-
13	-	50	58	31.18	10.53	1340	.89	.54	.18	.38	.13
	1	-		30.95	11.07	658	-	-	-	-	-
	5	-		31.40	9.97	682	-	-	-	-	-
14	-	50	60	32.99	11.45	1293	.90	.55	.19	.41	.12
	1	-		32.19	11.67	648	-	-	-	-	-
	2	-		33.80	11.16	645	-	-	-	-	-
15	-	50	60	33.65	10.34	1178	.88	.56	.17	.38	.12
	2	-		33.96	10.52	653	-	-	-	-	-
	6	-		33.26	10.08	525	-	-	-	-	-
16	-	50	58	34.99	10.31	1306	.89	.60	.18	.39	.12
	2	-		36.13	10.32	647	-	-	-	-	-
	3	-		33.88	10.17	659	-	-	-	-	-
17	-	50	57	34.25	10.58	1320	.90	.54	.19	.41	.11
	3	-		30.96	10.64	645	-	-	-	-	-
	5	-		31.14	10.51	675	-	-	-	-	-
18	-	50	59	34.25	10.74	1341	.90	.59	.19	.41	.12
	3	-		33.58	10.94	632	-	-	-	-	-
	4	-		34.84	10.51	709	-	-	-	-	-
19	-	50	60	34.63	11.24	1209	.90	.58	.19	.41	.11
	4	-		34.38	11.33	697	-	-	-	-	-
	6	-		34.97	11.10	512	-	-	-	-	-

[1] - Includes p-values for constructed-response items obtained by dividing the maximum number of points.

Table 21. HSPT in Science Pilot
Item Statistics by Form
Form 12

ITEM	TYPE*	N	P VAL	RAW MEAN	STDV
1	M	1361	0.82292	0.82292	0.38187
2	M	1361	0.37325	0.37325	0.48385
3	M	1361	0.48420	0.48420	0.49993
4	M	1361	0.69140	0.69140	0.46208
5	M	1361	0.84644	0.84644	0.36066
6	M	1361	0.38281	0.38281	0.48625
7	M	1361	0.55474	0.55474	0.49718
8	M	1361	0.86260	0.86260	0.34439
9	M	1361	0.44526	0.44526	0.49718
10	M	1361	0.29317	0.29317	0.45538
11	M	1361	0.64364	0.64364	0.47910
12	M	1361	0.75459	0.75459	0.43049
13	M	1361	0.92799	0.92799	0.25859
14	O	1361	0.51029	1.02057	0.65272
15	O	1361	0.72998	1.45996	0.73368
16	O	1361	0.65687	1.31374	0.78170
17	M	1361	0.20573	0.20573	0.40438
18	M	1361	0.37105	0.37105	0.48326
19	M	1361	0.56429	0.56429	0.49603
20	M	1361	0.28582	0.28582	0.45197
21	M	1361	0.40485	0.40485	0.49104
22	M	1361	0.73916	0.73916	0.43925
23	M	1361	0.62675	0.62675	0.48385
24	M	1361	0.80676	0.80676	0.39498
25	M	1361	0.58633	0.58633	0.49267
26	M	1361	0.84717	0.84717	0.35995
27	M	1361	0.65099	0.65099	0.47683
28	M	1361	0.66569	0.66569	0.47192
29	M	1361	0.39971	0.39971	0.49002
30	O	1361	0.23953	0.71859	0.92483
31	O	1361	0.60838	1.21675	0.72612
32	O	1361	0.65797	1.31594	0.75056
33	M	1361	0.67450	0.67450	0.46873
34	M	1361	0.72961	0.72961	0.44432
35	M	1361	0.86921	0.86921	0.33729
36	M	1361	0.75239	0.75239	0.43178
37	M	1361	0.57825	0.57825	0.49402
38	M	1361	0.75386	0.75386	0.43092
39	M	1361	0.78031	0.78031	0.41419
40	M	1361	0.66275	0.66275	0.47295
41	M	1361	0.65173	0.65173	0.47660
42	M	1361	0.82586	0.82586	0.37937
43	M	1361	0.56209	0.56209	0.49631
44	M	1361	0.68185	0.68185	0.46593
45	M	1361	0.55988	0.55988	0.49658
46	O	1361	0.43277	0.86554	0.83683
47	M	1361	0.50625	0.50625	0.50014
48	M	1361	0.53196	0.53196	0.49916
49	M	1361	0.20500	0.20500	0.40385
50	O	1361	0.50478	1.00955	0.88237

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 12

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SD_OE	MEAN_P	SD_P
TOTAL	1361	25.56280	7.20574	8.92065	4.27367	0.58447	0.18200
MALE	692	26.00000	7.72014	8.80347	4.60499	0.58989	0.19556
FEMALE	664	25.12650	6.60378	9.04518	3.90433	0.57918	0.16673
WHITE	1032	26.71510	6.68962	9.72093	3.84516	0.61756	0.16538
AF-AM	129	18.87600	6.44933	4.17829	3.74468	0.39075	0.15738

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1361	34.48350	10.73780	0.29106	0.40869	0.12889	0.37128	0.10264
MALE	692	34.80350	11.53820	0.43862	0.43454	0.12950	0.39684	0.10393
FEMALE	664	34.17170	9.83720	0.38176	0.37961	0.13806	0.34189	0.11198
WHITE	1032	36.43600	9.75720	0.30373	0.38443	0.11674	0.35274	0.09757
AF-AM	129	23.05430	9.28570	0.81756	0.35297	0.17710	0.30965	0.15427

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.77392	USING	S	12
0.63637	CONSTRUCTING	S	12
0.48681	REFLECTING	S	12
0.47024	LIFE	S	12
0.47607	PHYSICAL	S	12
0.70230	EARTH	S	12

MEAN_MC - Mean score of multiple-choice items
SD_MC - Standard deviation of multiple-choice items
MEAN_OE - Mean score of constructed-response items
SD_OE - Standard deviation of constructed-response items
MEAN_P - Mean p value of multiple-choice items
SD_P - Standard deviation of p value

MEAN_T - Mean score of total test
SD_T - Standard deviation of total test
STDERR_T - Standard error of the total test
MRITT - Mean item-total correlation of the test
SDRITT - Standard deviation of item-total correlation
MR_MC - Mean correlation of item vs. total MC items
SDR_MC - Standard of correlation of item vs. total MC items

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 13

ITEM	TYPE*	N	P_VAL	RAW MEAN	STDV
1	M	1340	0.60522	0.60522	0.48899
2	M	1340	0.43060	0.43060	0.49534
3	M	1340	0.86567	0.86567	0.34113
4	M	1340	0.66940	0.66940	0.47060
5	M	1340	0.55746	0.55746	0.49687
6	M	1340	0.67313	0.67313	0.46924
7	M	1340	0.38209	0.38209	0.48608
8	M	1340	0.72164	0.72164	0.44836
9	M	1340	0.41418	0.41418	0.49276
10	M	1340	0.83209	0.83209	0.37393
11	M	1340	0.87015	0.87015	0.33626
12	M	1340	0.49627	0.49627	0.50017
13	M	1340	0.43657	0.43657	0.49615
14	O	1340	0.38433	0.76866	0.88902
15	O	1340	0.56642	1.13284	0.70438
16	O	1340	0.50448	1.00896	0.69775
17	M	1340	0.61866	0.61866	0.48590
18	M	1340	0.54851	0.54851	0.49783
19	M	1340	0.80896	0.80896	0.39327
20	M	1340	0.33358	0.33358	0.47167
21	M	1340	0.51791	0.51791	0.49987
22	M	1340	0.75746	0.75746	0.42878
23	M	1340	0.45522	0.45522	0.49818
24	M	1340	0.64627	0.64627	0.47831
25	M	1340	0.74478	0.74478	0.43615
26	M	1340	0.77388	0.77388	0.41847
27	M	1340	0.40672	0.40672	0.49140
28	M	1340	0.62388	0.62388	0.48459
29	M	1340	0.39552	0.39552	0.48915
30	O	1340	0.53246	1.06493	0.87166
31	O	1340	0.52575	1.05149	0.92893
32	O	1340	0.23731	0.47463	0.73438
33	M	1340	0.64627	0.64627	0.47831
34	M	1340	0.40299	0.40299	0.49068
35	M	1340	0.25746	0.25746	0.43740
36	M	1340	0.60448	0.60448	0.48915
37	M	1340	0.62164	0.62164	0.48516
38	M	1340	0.18507	0.18507	0.38850
39	M	1340	0.62537	0.62537	0.48421
40	M	1340	0.48060	0.48060	0.49981
41	M	1340	0.43358	0.43358	0.49575
42	M	1340	0.35075	0.35075	0.47738
43	M	1340	0.40448	0.40448	0.49097
44	M	1340	0.54254	0.54254	0.49837
45	M	1340	0.62388	0.62388	0.48459
46	O	1340	0.47463	0.94925	0.83005
47	M	1340	0.54552	0.54552	0.49811
48	M	1340	0.42687	0.42687	0.49481
49	M	1340	0.58955	0.58955	0.49210
50	O	1340	0.70112	1.40224	0.86686

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 13

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SD_OE	MEAN_P	SD_P
TOTAL	1340	23.32690	7.09231	7.85299	4.17701	0.53758	0.18157
MALE	646	23.98300	7.48692	7.70588	4.23974	0.54636	0.18928
FEMALE	680	22.72790	6.58027	8.00000	4.08957	0.52979	0.17202
WHITE	1018	24.45680	6.74574	8.49214	3.91232	0.56808	0.17057
AF-AM	144	17.38190	5.12910	4.22222	3.81833	0.37249	0.14175

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1340	31.17990	10.53100	0.28768	0.38286	0.12740	0.34755	0.10432
MALE	646	31.68890	10.97820	0.43193	0.40131	0.12546	0.36810	0.10531
FEMALE	680	30.72790	9.97720	0.38261	0.36126	0.13582	0.32256	0.10951
WHITE	1018	32.94890	9.89280	0.31006	0.36755	0.12040	0.33567	0.10157
AF-AM	144	21.60420	8.22180	0.68515	0.29141	0.18835	0.23895	0.14883

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.76736	USING	S	13
0.54282	CONSTRUCTING	S	13
0.37099	REFLECTING	S	13
0.58678	LIFE	S	13
0.51240	PHYSICAL	S	13
0.54296	EARTH	S	13

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 14

ITEM	TYPE*	N	P VAL	RAW MEAN	STDV
1	M	1293	0.79118	0.79118	0.40662
2	M	1293	0.76643	0.76643	0.42326
3	M	1293	0.82057	0.82057	0.38386
4	M	1293	0.83527	0.83527	0.37108
5	M	1293	0.77417	0.77417	0.41829
6	M	1293	0.58159	0.58159	0.49349
7	M	1293	0.53132	0.53132	0.49921
8	M	1293	0.38051	0.38051	0.48570
9	M	1293	0.35886	0.35886	0.47985
10	M	1293	0.50271	0.50271	0.50019
11	M	1293	0.68368	0.68368	0.46522
12	M	1293	0.67672	0.67672	0.46791
13	M	1293	0.69760	0.69760	0.45947
14	O	1293	0.20727	0.41454	0.66990
15	O	1293	0.51315	1.02630	0.83883
16	O	1293	0.48763	0.97525	0.84440
17	M	1293	0.60402	0.60402	0.48925
18	M	1293	0.62954	0.62954	0.48311
19	M	1293	0.69374	0.69374	0.46112
20	M	1293	0.73859	0.73859	0.43957
21	M	1293	0.60093	0.60093	0.48990
22	M	1293	0.64965	0.64965	0.47726
23	M	1293	0.33797	0.33797	0.47320
24	M	1293	0.54911	0.54911	0.49777
25	M	1293	0.82831	0.82831	0.37726
26	M	1293	0.43078	0.43078	0.49538
27	M	1293	0.66744	0.66744	0.47131
28	M	1293	0.35576	0.35576	0.47893
29	M	1293	0.88708	0.88708	0.31661
30	O	1293	0.36736	1.10209	1.08198
31	O	1293	0.59474	1.18948	0.76321
32	O	1293	0.43078	0.86156	0.94845
33	M	1293	0.65971	0.65971	0.47399
34	M	1293	0.39056	-0.39056	0.48807
35	M	1293	0.37664	0.37664	0.48473
36	M	1293	0.68910	0.68910	0.46304
37	M	1293	0.71694	0.71694	0.45066
38	M	1293	0.43001	0.43001	0.49527
39	M	1293	0.54215	0.54215	0.49841
40	M	1293	0.42923	0.42923	0.49516
41	M	1293	0.55839	0.55839	0.49677
42	M	1293	0.54679	0.54679	0.49800
43	M	1293	0.65429	0.65429	0.47578
44	M	1293	0.60866	0.60866	0.48824
45	M	1293	0.55916	0.55916	0.49668
46	O	1293	0.26167	0.78500	0.93778
47	M	1293	0.59010	0.59010	0.49201
48	M	1293	0.48569	0.48569	0.49999
49	M	1293	0.71230	0.71230	0.45287
50	O	1293	0.65623	1.31245	0.74431

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 14

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SO_OE	MEAN_P	SD_P
TOTAL	1293	25.32330	7.63987	7.66667	4.72931	0.54983	0.19086
MALE	654	25.99540	8.21914	7.64067	4.94140	0.56060	0.20384
FEMALE	628	24.67680	6.90438	7.72771	4.50809	0.54007	0.17553
WHITE	1026	25.95130	7.41005	8.09844	4.57022	0.56750	0.18380
AF-AM	89	19.48310	6.75229	3.08989	3.66073	0.37622	0.15795

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1293	32.98990	11.45170	0.31847	0.41020	0.12143	0.37410	0.08811
MALE	654	33.63610	12.23030	0.47824	0.43954	0.11839	0.40614	0.08943
FEMALE	628	32.40450	10.53190	0.42027	0.37578	0.13397	0.33554	0.09827
WHITE	1026	34.04970	11.02810	0.34429	0.39926	0.11601	0.36532	0.08394
AF-AM	89	22.57300	9.47710	1.00457	0.35529	0.16804	0.31472	0.14665

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.78343	USING	S	14
0.68854	CONSTRUCTING	S	14
0.42880	REFLECTING	S	14
0.56111	LIFE	S	14
0.54448	PHYSICAL	S	14
0.59578	EARTH	S	14

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 15

ITEM	TYPE*	N	P_VAL	RAW MEAN	STDV
1	M	1178	0.48472	0.48472	0.49998
2	M	1178	0.86248	0.86248	0.34454
3	M	1178	0.94143	0.94143	0.23493
4	M	1178	0.27674	0.27674	0.44758
5	M	1178	0.34211	0.34211	0.47462
6	M	1178	0.58404	0.58404	0.49310
7	M	1178	0.73599	0.73599	0.44099
8	M	1178	0.72750	0.72750	0.44543
9	M	1178	0.83192	0.83192	0.37410
10	M	1178	0.74533	0.74533	0.43586
11	M	1178	0.50000	0.50000	0.50021
12	M	1178	0.50424	0.50424	0.50019
13	M	1178	0.70543	0.70543	0.45604
14	O	1178	0.47963	1.43888	0.90317
15	O	1178	0.47496	0.94992	0.64788
16	O	1178	0.48854	0.97708	0.67883
17	M	1178	0.78778	0.78778	0.40906
18	M	1178	0.32598	0.32598	0.46894
19	M	1178	0.59932	0.59932	0.49024
20	M	1178	0.38200	0.38200	0.48608
21	M	1178	0.46435	0.46435	0.49894
22	M	1178	0.64601	0.64601	0.47841
23	M	1178	0.37267	0.37267	0.48372
24	M	1178	0.62309	0.62309	0.48482
25	M	1178	0.92954	0.92954	0.25603
26	M	1178	0.81664	0.81664	0.38713
27	M	1178	0.24024	0.24024	0.42741
28	M	1178	0.38370	0.38370	0.48649
29	M	1178	0.57470	0.57470	0.49460
30	O	1178	0.61602	1.84805	1.12808
31	O	1178	0.42997	0.85993	0.78424
32	O	1178	0.59380	1.18761	0.85456
33	M	1178	0.64771	0.64771	0.47789
34	M	1178	0.56367	0.56367	0.49614
35	M	1178	0.75806	0.75806	0.42844
36	M	1178	0.56112	0.56112	0.49646
37	M	1178	0.42020	0.42020	0.49380
38	M	1178	0.48557	0.48557	0.50000
39	M	1178	0.64007	0.64007	0.48018
40	M	1178	0.62818	0.62818	0.48350
41	M	1178	0.57046	0.57046	0.49522
42	M	1178	0.64941	0.64941	0.47736
43	M	1178	0.80815	0.80815	0.39392
44	M	1178	0.55178	0.55178	0.49752
45	M	1178	0.50000	0.50000	0.50021
46	O	1178	0.59805	1.19610	0.78846
47	M	1178	0.55857	0.55857	0.49677
48	M	1178	0.55688	0.55688	0.49697
49	M	1178	0.58574	0.58574	0.49280
50	O	1178	0.15747	0.31494	0.59462

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 15

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SD_OE	MEAN_P	SD_P
TOTAL	1178	24.87350	6.98695	8.77250	4.20068	0.56077	0.17227
MALE	579	25.57510	7.48464	8.70466	4.46914	0.57133	0.18432
FEMALE	591	24.18100	6.39142	8.83926	3.93415	0.55034	0.15923
WHITE	987	25.44070	6.76970	9.04863	4.05661	0.57482	0.16544
AF-AM	43	17.74420	5.32779	4.86047	4.48055	0.37674	0.15180

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SORITT	MR_MC	SDR_MC
TOTAL	1178	33.64600	10.33620	0.30115	0.38265	0.11796	0.34755	0.09060
MALE	579	34.27980	11.05900	0.45960	0.40815	0.11157	0.37447	0.08415
FEMALE	591	33.02030	9.55380	0.39299	0.35548	0.13598	0.31772	0.11086
WHITE	987	34.48940	9.92660	0.31597	0.37171	0.11585	0.33844	0.09239
AF-AM	43	22.60470	9.10820	1.38899	0.31953	0.22068	0.25209	0.15919

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.73554	USING	S	15
0.62238	CONSTRUCTING	S	15
0.47649	REFLECTING	S	15
0.53524	LIFE	S	15
0.47386	PHYSICAL	S	15
0.4909	EARTH	S	15

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 16

ITEM	TYPE*	N	P_VAL	RAW MEAN	STDV
1	M	1306	0.47933	0.47933	0.49976
2	M	1306	0.64548	0.64548	0.47855
3	M	1306	0.84456	0.84456	0.36246
4	M	1306	0.55054	0.55054	0.49763
5	M	1306	0.48928	0.48928	0.50008
6	M	1306	0.56662	0.56662	0.49573
7	M	1306	0.69449	0.69449	0.46080
8	M	1306	0.70444	0.70444	0.45647
9	M	1306	0.94334	0.94334	0.23128
10	M	1306	0.49005	0.49005	0.50009
11	M	1306	0.68606	0.68606	0.46427
12	M	1306	0.75421	0.75421	0.43072
13	M	1306	0.77871	0.77871	0.41527
14	O	1306	0.85911	1.71822	0.53841
15	O	1306	0.45904	0.91807	0.81190
16	O	1306	0.45061	0.90123	0.81286
17	M	1306	0.75804	0.75804	0.42843
18	M	1306	0.64012	0.64012	0.48015
19	M	1306	0.59648	0.59648	0.49079
20	M	1306	0.46861	0.46861	0.49920
21	M	1306	0.66233	0.66233	0.47310
22	M	1306	0.37902	0.37902	0.48533
23	M	1306	0.58040	0.58040	0.49368
24	M	1306	0.63247	0.63247	0.48232
25	M	1306	0.77642	0.77642	0.41681
26	M	1306	0.71210	0.71210	0.45296
27	M	1306	0.70214	0.70214	0.45749
28	M	1306	0.67075	0.67075	0.47012
29	M	1306	0.21057	0.21057	0.40787
30	O	1306	0.39855	0.79709	0.82825
31	O	1306	0.65773	1.31547	0.65859
32	O	1306	0.57312	1.14625	0.70416
33	M	1306	0.49770	0.49770	0.50019
34	M	1306	0.44028	0.44028	0.49661
35	M	1306	0.48086	0.48086	0.49982
36	M	1306	0.58959	0.58959	0.49210
37	M	1306	0.51761	0.51761	0.49988
38	M	1306	0.51608	0.51608	0.49993
39	M	1306	0.54977	0.54977	0.49771
40	M	1306	0.73354	0.73354	0.44228
41	M	1306	0.79403	0.79403	0.40457
42	M	1306	0.48469	0.48469	0.49996
43	M	1306	0.85222	0.85222	0.35502
44	M	1306	0.35528	0.35528	0.47878
45	M	1306	0.80092	0.80092	0.39946
46	O	1306	0.45674	0.91348	0.90262
47	M	1306	0.87519	0.87519	0.33063
48	M	1306	0.53446	0.53446	0.49900
49	M	1306	0.86141	0.86141	0.34565
50	O	1306	0.49158	0.98315	0.85752

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 16

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SD_OE	MEAN_P	SD_P
TOTAL	1306	26.30020	7.16731	8.69296	3.94074	0.60333	0.17779
MALE	633	26.70140	7.62414	8.69984	4.19406	0.61037	0.19009
FEMALE	663	25.92160	6.68112	8.69985	3.66729	0.59692	0.16478
WHITE	1060	26.83580	6.97828	8.98019	3.85607	0.61752	0.17260
AF-AM	83	22.65060	6.69990	6.69880	3.24125	0.50602	0.15901

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1306	34.99310	10.31190	0.28534	0.39280	0.11706	0.36271	0.09921
MALE	633	35.40130	11.02530	0.43822	0.41791	0.12150	0.38679	0.10454
FEMALE	663	34.62140	9.55700	0.37116	0.36686	0.11956	0.33749	0.10266
WHITE	1060	35.81600	10.01090	0.30748	0.38533	0.11552	0.35624	0.09874
AF-AM	83	29.34940	9.22280	1.01233	0.34890	0.14567	0.32331	0.14087

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.76829	USING	S	16
0.57958	CONSTRUCTING	S	16
0.46614	REFLECTING	S	16
0.57515	LIFE	S	16
0.53238	PHYSICAL	S	16
0.47016	EARTH	S	16

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 17

ITEM	TYPE*	N	P_VAL	RAW MEAN	STDV
1	M	1320	0.61439	0.61439	0.48692
2	M	1320	0.74470	0.74470	0.43620
3	M	1320	0.59773	0.59773	0.49054
4	M	1320	0.74318	0.74318	0.43704
5	M	1320	0.58561	0.58561	0.49280
6	M	1320	0.37045	0.37045	0.48311
7	M	1320	0.57955	0.57955	0.49382
8	M	1320	0.89697	0.89697	0.30411
9	M	1320	0.66894	0.66894	0.47077
10	M	1320	0.70076	0.70076	0.45810
11	M	1320	0.52727	0.52727	0.49944
12	M	1320	0.58788	0.58788	0.49240
13	M	1320	0.61212	0.61212	0.48745
14	O	1320	0.59773	1.19545	0.82492
15	O	1320	0.60076	1.20152	0.71505
16	O	1320	0.57500	1.15000	0.72028
17	M	1320	0.59545	0.59545	0.49099
18	M	1320	0.54167	0.54167	0.49845
19	M	1320	0.65606	0.65606	0.47520
20	M	1320	0.41742	0.41742	0.49332
21	M	1320	0.58864	0.58864	0.49227
22	M	1320	0.38712	0.38712	0.48728
23	M	1320	0.59167	0.59167	0.49171
24	M	1320	0.60909	0.60909	0.48814
25	M	1320	0.54773	0.54773	0.49791
26	M	1320	0.51212	0.51212	0.50004
27	M	1320	0.70606	0.70606	0.45574
28	M	1320	0.89091	0.89091	0.31187
29	M	1320	0.54318	0.54318	0.49832
30	O	1320	0.24697	0.49394	0.69546
31	O	1320	0.18864	0.37727	0.64228
32	O	1320	0.32955	0.32955	0.47023
33	M	1320	0.52879	0.52879	0.49936
34	M	1320	0.55303	0.55303	0.49737
35	M	1320	0.88788	0.88788	0.31564
36	M	1320	0.62500	0.62500	0.48431
37	M	1320	0.70909	0.70909	0.45435
38	M	1320	0.52045	0.52045	0.49977
39	M	1320	0.61136	0.61136	0.48763
40	M	1320	0.57955	0.57955	0.49382
41	M	1320	0.58561	0.58561	0.49280
42	M	1320	0.52576	0.52576	0.49953
43	M	1320	0.54167	0.54167	0.49845
44	M	1320	0.49848	0.49848	0.50019
45	M	1320	0.52803	0.52803	0.49940
46	O	1320	0.23636	0.47273	0.70147
47	M	1320	0.84924	0.84924	0.35795
48	M	1320	0.46212	0.46212	0.49875
49	M	1320	0.55530	0.55530	0.49712
50	O	1320	0.22576	0.45152	0.62601

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 17

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SO_OE	MEAN_P	SD_P
TOTAL	1320	25.37800	7.98569	5.67197	3.43453	0.54474	0.18564
MALE	648	25.39510	8.49100	5.08951	3.42795	0.53482	0.19491
FEMALE	658	25.38150	7.49746	6.25228	3.34521	0.55498	0.17629
WHITE	1031	26.18140	7.81268	5.99224	3.39445	0.56445	0.18146
AF-AM	142	20.32390	7.22343	3.67606	3.07483	0.42105	0.16608

GROUP	N	MEAN_T	SD_T	STOERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1320	31.05000	10.58150	0.29125	0.41102	0.10508	0.39109	0.09035
MALE	648	30.48460	11.10980	0.43643	0.43244	0.10240	0.41529	0.08705
FEMALE	658	31.63370	10.04830	0.39173	0.39090	0.11813	0.36862	0.10683
WHITE	1031	32.17360	10.34320	0.32213	0.40533	0.10392	0.38616	0.09136
AF-AM	142	24.00000	9.46660	0.79442	0.37068	0.12047	0.34584	0.10237

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.80176	USING	S	17
0.66419	CONSTRUCTING	S	17
0.47477	REFLECTING	S	17
0.59102	LIFE	S	17
0.54850	PHYSICAL	S	17
0.62215	EARTH	S	17

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 18

ITEM	TYPE*	N	P VAL	RAW MEAN	STDV
1	M	1341	0.95749	0.95749	0.20181
2	M	1341	0.52871	0.52871	0.49936
3	M	1341	0.82327	0.82327	0.38159
4	M	1341	0.66145	0.66145	0.47339
5	M	1341	0.36913	0.36913	0.48275
6	M	1341	0.47278	0.47278	0.49944
7	M	1341	0.65697	0.65697	0.47490
8	M	1341	0.37360	0.37360	0.48394
9	M	1341	0.81133	0.81133	0.39139
10	M	1341	0.58240	0.58240	0.49335
11	M	1341	0.50112	0.50112	0.50019
12	M	1341	0.78151	0.78151	0.41338
13	M	1341	0.87174	0.87174	0.33451
14	O	1341	0.74944	1.49888	0.76258
15	O	1341	0.63609	1.27218	0.80560
16	O	1341	0.61782	1.23565	0.82928
17	M	1341	0.69053	0.69053	0.46245
18	M	1341	0.56600	0.56600	0.49581
19	M	1341	0.29605	0.29605	0.45668
20	M	1341	0.73304	0.73304	0.44254
21	M	1341	0.60477	0.60477	0.48908
22	M	1341	0.70395	0.70395	0.45668
23	M	1341	0.42878	0.42878	0.49509
24	M	1341	0.74944	0.74944	0.43350
25	M	1341	0.78896	0.78896	0.40820
26	M	1341	0.70917	0.70917	0.45431
27	M	1341	0.72185	0.72185	0.44826
28	M	1341	0.67562	0.67562	0.46832
29	M	1341	0.65250	0.65250	0.47635
30	O	1341	0.40455	0.80910	0.81823
31	O	1341	0.31954	0.63908	0.67321
32	O	1341	0.48471	0.96943	0.75000
33	M	1341	0.60626	0.60626	0.48876
34	M	1341	0.52871	0.52871	0.49936
35	M	1341	0.52871	0.52871	0.49936
36	M	1341	0.86801	0.86801	0.33861
37	M	1341	0.44892	0.44892	0.49757
38	M	1341	0.68307	0.68307	0.46545
39	M	1341	0.74124	0.74124	0.43812
40	M	1341	0.28784	0.28784	0.45293
41	M	1341	0.57867	0.57867	0.49396
42	M	1341	0.74198	0.74198	0.43771
43	M	1341	0.28635	0.28635	0.45223
44	M	1341	0.71141	0.71141	0.45328
45	M	1341	0.81133	0.81133	0.39139
46	O	1341	0.27442	0.54884	0.68788
47	M	1341	0.59284	0.59284	0.49149
48	M	1341	0.75391	0.75391	0.43089
49	M	1341	0.74720	0.74720	0.43478
50	O	1341	0.32364	0.64728	0.81870

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 18

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SD_DE	MEAN P	SD_P
TOTAL	1341	26.62860	7.41384	7.62043	4.04433	0.59050	0.18516
MALE	678	27.36280	8.00800	7.47493	4.13957	0.60065	0.19713
FEMALE	651	25.84020	6.68217	7.74962	3.93215	0.57914	0.17158
WHITE	1067	27.58480	6.99913	8.18369	3.84088	0.61670	0.17359
AF-AM	115	20.13040	7.37174	4.09565	3.53175	0.41769	0.17679

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1341	34.24910	10.73910	0.29326	0.41159	0.12258	0.38027	0.10182
MALE	678	34.83780	11.43360	0.43910	0.44035	0.12004	0.41222	0.10453
FEMALE	651	33.58990	9.95170	0.39004	0.37957	0.13775	0.34333	0.11276
WHITE	1067	35.76850	10.06830	0.30823	0.39497	0.11512	0.36585	0.09535
AF-AM	115	24.22610	10.25380	0.95618	0.38308	0.15271	0.35030	0.13812

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.80723	USING	S	18
0.53443	CONSTRUCTING	S	18
0.51522	REFLECTING	S	18
0.56158	LIFE	S	18
0.54253	PHYSICAL	S	18
0.66907	EARTH	S	18

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 19

ITEM	TYPE*	N	P VAL	RAW MEAN	STDV
1	M	1209	0.75848	0.75848	0.42818
2	M	1209	0.47808	0.47808	0.49973
3	M	1209	0.53184	0.53184	0.49919
4	M	1209	0.55914	0.55914	0.49670
5	M	1209	0.50207	0.50207	0.50020
6	M	1209	0.88337	0.88337	0.32111
7	M	1209	0.39620	0.39620	0.48931
8	M	1209	0.49628	0.49628	0.50019
9	M	1209	0.34905	0.34905	0.47687
10	M	1209	0.73945	0.73945	0.43911
11	M	1209	0.64764	0.64764	0.47790
12	M	1209	0.77667	0.77667	0.41665
13	M	1209	0.60050	0.60050	0.49000
14	O	1209	0.76592	2.29777	1.01349
15	O	1209	0.80025	1.60050	0.65738
16	O	1209	0.66046	1.32093	0.85326
17	M	1209	0.38710	0.38710	0.48729
18	M	1209	0.39041	0.39041	0.48804
19	M	1209	0.38213	0.38213	0.48611
20	M	1209	0.64433	0.64433	0.47891
21	M	1209	0.68983	0.68983	0.46276
22	M	1209	0.45079	0.45079	0.49778
23	M	1209	0.44500	0.44500	0.49717
24	M	1209	0.82713	0.82713	0.37829
25	M	1209	0.59967	0.59967	0.49017
26	M	1209	0.59967	0.59967	0.49017
27	M	1209	0.66998	0.66998	0.47042
28	M	1209	0.67825	0.67825	0.46734
29	M	1209	0.35732	0.35732	0.47941
30	O	1209	0.35153	1.05459	0.94044
31	O	1209	0.61497	1.22994	0.84673
32	O	1209	0.35608	0.71216	0.69995
33	M	1209	0.46071	0.46071	0.49866
34	M	1209	0.81638	0.81638	0.38734
35	M	1209	0.80397	0.80397	0.39716
36	M	1209	0.87097	0.87097	0.33537
37	M	1209	0.32258	0.32258	0.46766
38	M	1209	0.35732	0.35732	0.47941
39	M	1209	0.73863	0.73863	0.43956
40	M	1209	0.62035	0.62035	0.48550
41	M	1209	0.81969	0.81969	0.38461
42	M	1209	0.51447	0.51447	0.50000
43	M	1209	0.82630	0.82630	0.37901
44	M	1209	0.65509	0.65509	0.47554
45	M	1209	0.79156	0.79156	0.40636
46	O	1209	0.37634	0.75269	0.81737
47	M	1209	0.58561	0.58561	0.49282
48	M	1209	0.38296	0.38296	0.48631
49	M	1209	0.51613	0.51613	0.49995
50	O	1209	0.36807	0.73615	0.90566

* M = Multiple-Choice Item, O = Constructed-Response Item

Table 21 (cont). HSPT in Science Pilot
Item Statistics by Form
Form 19

GROUP	N	MEAN_MC	SD_MC	MEAN_OE	SD_OE	MEAN_P	SD_P
TOTAL	1209	24.92310	7.55344	9.70470	4.45280	0.57713	0.18739
MALE	589	25.84890	7.86575	9.44140	4.56719	0.58817	0.19426
FEMALE	614	24.05860	7.10787	10.00980	4.30257	0.56781	0.17979
WHITE	991	25.76990	7.28549	10.19480	4.27412	0.59941	0.17946
AF-AM	86	16.65120	5.56410	4.93020	3.98170	0.35969	0.14587

GROUP	N	MEAN_T	SD_T	STDERR_T	MRITT	SDRITT	MR_MC	SDR_MC
TOTAL	1209	34.62780	11.24330	0.32335	0.41202	0.11195	0.37878	0.08283
MALE	589	35.29030	11.65590	0.48027	0.42930	0.10671	0.39871	0.08193
FEMALE	614	34.06840	10.78770	0.43535	0.39426	0.12443	0.35656	0.09040
WHITE	991	35.96470	10.76770	0.34205	0.40063	0.10590	0.36940	0.07794
AF-AM	86	21.58140	8.75210	0.94376	0.31096	0.19335	0.25818	0.15779

Alpha Coefficients For Science Subscales

ALPHA	SUBSCL	CONTENT	FORM
0.80866	USING	S	19
0.56532	CONSTRUCTING	S	19
0.45838	REFLECTING	S	19
0.60469	LIFE	S	19
0.54185	PHYSICAL	S	19
0.65176	EARTH	S	19

Table 23. HSPT in Science Pilot
Mean Interrater Agreement Based on First Two Readers

One-point item (0-1)

agree	adjacent
82.0%	18.0%

form 17 #6

Two-point items (0-2)

agree	adjacent	nonadjacent
72.4%	25.3%	2.3%

form 12 #1
form 12 #2
form 12 #3
form 12 #5
form 12 #6
form 12 #7
form 13 #1
form 13 #2
form 13 #3
form 13 #4
form 13 #5
form 13 #6
form 13 #7
form 13 #8
form 14 #1
form 14 #2
form 14 #3
form 14 #5
form 14 #6
form 14 #8
form 15 #2
form 15 #3
form 15 #5
form 15 #6
form 15 #7
form 15 #8
form 16 #1
form 16 #2
form 16 #3
form 16 #4
form 16 #5
form 16 #6
form 16 #7
form 16 #8
form 17 #1
form 17 #2
form 17 #3
form 17 #4
form 17 #5
form 17 #7
form 17 #8
form 18 #1
form 18 #2
form 18 #3
form 18 #5
form 18 #6
form 18 #7
form 18 #8
form 19 #2
form 19 #3
form 19 #5
form 19 #6
form 19 #7
form 19 #8

Three-point items (0-3)

agree	adjacent	nonadjacent
69.2%	27.5%	3.3%

form 12 #4
form 14 #4
form 14 #7
form 15 #1
form 15 #4
form 18 #4
form 19 #1
form 19 #4

Table 24. HSPT in Science Pilot
Interrater Agreement by Item

Agreement between first 2 readers:

1 = agree
2 = adjacent
3 = nonadjacent
. = student's response invalid

Form 12

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	120	.	.	.
1	918	71.9	918	71.9
2	354	27.7	1272	99.6
3	5	0.4	1277	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	148	.	.	.
1	872	69.8	872	69.8
2	363	29.1	1235	98.9
3	14	1.1	1249	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	182	.	.	.
1	803	66.1	803	66.1
2	384	31.6	1187	97.7
3	28	2.3	1215	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	286	.	.	.
1	710	63.9	710	63.9
2	322	29.0	1032	92.9
3	79	7.1	1111	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	188	.	.	.
1	828	68.5	828	68.5
2	349	28.9	1177	97.4
3	32	2.6	1209	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	166	.	.	.
1	784	63.7	784	63.7
2	419	34.0	1203	97.7
3	28	2.3	1231	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	265	.	.	.
1	834	73.7	834	73.7
2	278	24.6	1112	98.2
3	20	1.8	1132	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	282	.	.	.
1	765	68.6	765	68.6
2	329	29.5	1094	98.1
3	21	1.9	1115	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 13

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	99	.	.	.
1	986	78.2	986	78.2
2	229	18.2	1215	96.4
3	46	3.6	1261	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	127	.	.	.
1	882	71.5	882	71.5
2	345	28.0	1227	99.5
3	6	0.5	1233	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	167	.	.	.
1	846	70.9	846	70.9
2	344	28.8	1190	99.7
3	3	0.3	1193	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	108	.	.	.
1	1007	80.4	1007	80.4
2	209	16.7	1216	97.1
3	36	2.9	1252	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	142	.	.	.
1	865	71.0	865	71.0
2	290	23.8	1155	94.8
3	63	5.2	1218	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	200	.	.	.
1	905	78.0	905	78.0
2	227	19.6	1132	97.6
3	28	2.4	1160	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	183	.	.	.
1	788	66.9	788	66.9
2	360	30.6	1148	97.5
3	29	2.5	1177	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	152	.	.	.
1	985	81.5	985	81.5
2	167	13.8	1152	95.4
3	56	4.6	1208	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 14

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	144	.	.	.
1	891	76.3	891	76.3
2	252	21.6	1143	97.9
3	25	2.1	1168	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	180	.	.	.
1	891	78.7	891	78.7
2	232	20.5	1123	99.2
3	9	0.8	1132	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	205	.	.	.
1	858	77.5	858	77.5
2	240	21.7	1098	99.2
3	9	0.8	1107	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	181	.	.	.
1	831	73.5	831	73.5
2	259	22.9	1090	96.4
3	41	3.6	1131	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	137	.	.	.
1	881	75.0	881	75.0
2	287	24.4	1168	99.4
3	7	0.6	1175	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	208	.	.	.
1	793	71.8	793	71.8
2	181	16.4	974	88.2
3	130	11.8	1104	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	372	.	.	.
1	674	71.7	674	71.7
2	253	26.9	927	98.6
3	13	1.4	940	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	177	.	.	.
1	792	69.0	783	69.0
2	341	30.0	1124	99.0
3	11	1.0	1135	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 15

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	55	.	.	.
1	833	71.3	833	71.3
2	309	26.4	1142	97.7
3	27	2.3	1169	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	153	.	.	.
1	824	76.9	824	76.9
2	229	21.4	1053	98.3
3	18	1.7	1071	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	184	.	.	.
1	766	73.7	766	73.7
2	262	25.2	1028	98.8
3	12	1.2	1040	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	150	.	.	.
1	700	65.2	700	65.2
2	320	29.8	1020	95.0
3	54	5.0	1074	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	179	.	.	.
1	718	68.7	718	68.7
2	297	28.4	1015	97.1
3	30	2.9	1045	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	218	.	.	.
1	596	59.2	596	59.2
2	370	36.8	966	96.0
3	40	4.0	1006	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	170	.	.	.
1	722	68.5	722	68.5
2	320	30.4	40	98.9
3	12	1.1	1054	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	350	.	.	.
1	666	76.2	666	76.2
2	183	20.9	849	97.1
3	25	2.9	874	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 16

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	34	.	.	.
1	911	71.1	911	71.1
2	365	28.5	1276	99.6
3	5	0.4	1281	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	108	.	.	.
1	748	62.0	748	62.0
2	386	32.0	1134	94.0
3	73	6.0	1207	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	145	.	.	.
1	699	59.7	699	59.7
2	405	34.6	1104	94.4
3	66	5.6	1170	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	113	.	.	.
1	767	63.8	767	63.8
2	368	30.6	1135	94.4
3	67	5.6	1202	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	73	.	.	.
1	924	74.4	924	74.4
2	309	24.9	1233	99.3
3	9	0.7	1242	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	71	.	.	.
1	730	58.7	730	58.7
2	484	38.9	1214	97.6
3	30	2.4	1244	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	205	.	.	.
1	782	70.5	782	70.5
2	293	26.4	1075	96.8
3	35	3.2	1110	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	157	.	.	.
1	743	64.2	743	64.2
2	355	30.7	1098	94.8
3	60	5.2	1158	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 17

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	104	.	.	.
1	940	75.8	940	75.8
2	286	23.1	1226	98.9
3	14	1.1	1240	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	106	.	.	.
1	972	78.5	972	78.5
2	264	21.3	1236	99.8
3	2	0.2	1238	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	129	.	.	.
1	950	78.2	950	78.2
2	261	21.5	1211	99.7
3	4	0.3	1215	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	163	.	.	.
1	813	68.8	813	68.8
2	316	26.8	1129	95.6
3	52	4.4	1181	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	142	.	.	.
1	989	82.3	969	82.3
2	197	16.4	1186	98.7
3	16	1.3	1202	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	185	.	.	.
1	950	82.0	950	82.0
2	201	18.0	1159	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	260	.	.	.
1	772	71.2	772	71.2
2	276	25.5	1048	96.7
3	36	3.3	1084	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	266	.	.	.
1	769	71.3	769	71.3
2	296	27.5	1065	98.8
3	13	1.2	1078	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 18

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	39	.	.	.
1	1058	79.7	1058	79.7
2	244	18.4	1302	98.0
3	26	2.0	1328	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	115	.	.	.
1	1037	82.8	1037	82.8
2	214	17.1	1251	99.9
3	1	0.1	1252	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	152	.	.	.
1	981	80.7	981	80.7
2	225	18.5	1206	99.3
3	9	0.7	1215	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	186	.	.	.
1	858	72.7	858	72.7
2	299	25.3	1157	98.0
3	24	2.0	1181	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	110	.	.	.
1	924	73.5	924	73.5
2	321	25.5	1245	99.0
3	12	1.0	1257	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	175	.	.	.
1	867	72.7	867	72.7
2	308	25.8	1175	98.6
3	17	1.4	1192	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	175	.	.	.
1	823	69.0	823	69.0
2	338	28.3	1161	97.3
3	32	2.7	1193	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	204	.	.	.
1	962	82.7	962	82.7
2	198	17.0	1160	99.7
3	3	0.3	1163	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Form 19

Constructed-Response 1

ITEM14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	63	.	.	.
1	866	71.9	866	71.9
2	313	26.0	1179	97.8
3	26	2.2	1205	100.0

Constructed-Response 2

ITEM15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	63	.	.	.
1	925	76.8	925	76.8
2	267	22.2	1192	98.9
3	13	1.1	1205	100.0

Constructed-Response 3

ITEM16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	111	.	.	.
1	761	65.8	761	65.8
2	335	29.0	1096	94.7
3	61	5.3	1157	100.0

Constructed-Response 4

ITEM30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	198	.	.	.
1	674	63.0	674	63.0
2	367	34.3	1041	97.3
3	29	2.7	1070	100.0

Table 24 (cont.). HSPT in Science Pilot
Interrater Agreement by Item

Constructed-Response 5

ITEM31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	191	.	.	.
1	831	77.2	831	77.2
2	235	21.8	1066	99.0
3	11	1.0	1077	100.0

Constructed-Response 6

ITEM32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	290	.	.	.
1	581	59.4	581	59.4
2	369	37.7	950	97.1
3	28	2.9	978	100.0

Constructed-Response 7

ITEM46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	216	.	.	.
1	818	77.8	818	77.8
2	221	21.0	1039	98.8
3	13	1.2	1052	100.0

Constructed-Response 8

ITEM50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
invalid	219	.	.	.
1	818	78.0	818	78.0
2	193	18.4	1011	96.4
3	38	3.6	1049	100.0

Table 25. HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 12

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	122	8.8	122	8.8
0	174	12.5	296	21.3
1	791	56.8	1087	78.1
2	305	21.9	1392	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	159	11.4	159	11.4
0	63	4.5	222	15.9
1	342	24.6	564	40.5
2	828	59.5	1392	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	207	14.9	207	14.9
0	89	6.4	296	21.3
1	397	28.5	693	49.8
2	699	50.2	1392	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	283	20.3	283	20.3
0	471	33.8	754	54.2
1	400	28.7	1154	82.9
2	135	9.7	1289	92.6
3	103	7.4	1392	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	217	15.6	217	15.6
0	55	4.0	272	19.5
1	580	41.7	852	61.2
2	540	38.8	1392	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	191	13.7	191	13.7
0	76	5.5	267	19.2
1	459	33.0	726	52.2
2	666	47.8	1392	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	282	20.3	282	20.3
0	329	23.6	611	43.9
1	384	27.6	995	71.5
2	397	28.5	1392	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	300	21.6	300	21.6
0	254	18.2	554	39.8
1	302	21.7	856	61.5
2	536	38.5	1392	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 13

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	145	10.7	145	10.7
0	591	43.5	736	54.1
1	213	15.7	949	69.8
2	411	30.2	1360	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	133	9.8	133	9.8
0	137	10.1	270	19.9
1	656	48.2	923	68.1
2	434	31.9	1360	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	170	12.5	170	12.5
0	166	12.2	336	24.7
1	691	50.8	1027	75.5
2	333	24.5	1360	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	144	10.6	144	10.6
0	342	25.1	486	35.7
1	319	23.5	805	59.2
2	555	40.8	1360	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	502	15.1	205	15.1
0	358	26.3	536	41.4
1	183	13.5	746	54.9
2	614	45.1	1360	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	228	16.8	228	16.8
0	690	50.7	918	67.5
1	248	18.2	1166	85.7
2	194	14.3	1360	100.0

Constructed-Response 7

Raw46	Frequency	Perc15.6ent	Cumulative Frequency	Cumulative Percent
.	212	15.6	212	15.6
0	305	22.4	517	38.0
1	414	30.4	931	68.5
2	429	31.5	1360	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	208	15.3	208	15.3
0	153	11.3	361	26.5
1	118	8.7	479	35.2
2	881	64.8	1360	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 14

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	169	12.9	169	12.9
0	740	56.4	909	69.3
1	270	20.6	1179	89.9
2	133	10.1	1312	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	189	14.4	189	14.4
0	267	20.4	456	34.8
1	384	29.3	810	64.0
2	472	36.0	1312	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	214	16.3	214	16.3
0	281	21.4	495	37.7
1	372	28.4	867	66.1
2	445	33.9	1312	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	181	13.8	181	13.8
0	377	28.7	558	42.5
1	238	18.1	796	60.7
2	359	27.4	1155	88.0
3	157	12.0	1312	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	144	11.0	144	11.0
0	151	11.5	295	22.5
1	495	37.7	790	60.2
2	522	39.8	1312	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	338	25.8	338	25.8
0	364	27.7	702	53.5
1	106	8.1	808	61.6
2	504	38.4	1312	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	372	28.4	372	28.4
0	290	22.1	662	50.5
1	379	28.9	1041	79.3
2	177	13.5	1218	92.8
3	94	7.2	1312	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	188	14.3	188	14.3
0	49	3.7	237	18.1
1	451	34.4	688	52.4
2	624	47.6	1312	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 15

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	55	4.5	55	4.5
0	82	6.8	137	11.3
1	629	52.0	766	63.4
2	235	19.4	1001	82.8
3	208	17.2	1209	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	166	13.7	166	13.7
0	137	11.3	303	25.1
1	685	56.7	988	81.7
2	221	18.3	1209	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	191	15.8	191	15.8
0	120	9.9	311	25.7
1	639	52.9	950	78.6
2	259	21.4	1209	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	146	12.1	146	12.1
0	89	7.4	235	19.4
1	222	18.4	457	37.8
2	287	23.7	744	61.5
3	465	38.5	1209	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	206	17.0	206	17.0
0	280	23.2	486	40.2
1	432	35.7	918	75.9
2	291	24.1	1209	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	254	21.0	254	21.0
0	116	9.6	370	30.6
1	27	22.9	647	53.5
2	562	46.5	1209	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	178	14.7	178	14.7
0	126	10.4	304	25.1
1	401	33.2	705	58.3
2	504	41.7	1209	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	370	30.6	370	30.6
0	549	45.4	919	76.0
1	209	17.3	1128	93.3
2	81	6.7	1209	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 16

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	39	3..1	39	3.0
0	19	1.4	58	4.4
1	257	19.5	315	24.0
2	1000	76.0	1315	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	181	13.8	181	13.8
0	312	23.7	493	37.5
1	440	33.5	933	71.0
2	382	29.0	1315	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	211	16.0	211	16.0
0	295	22.4	506	38.5
1	434	33.0	940	71.5
2	375	28.5	1315	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	180	13.7	180	13.7
0	434	33.0	614	46.7
1	359	27.3	973	74.0
2	342	26.0	1315	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	82	6.2	82	6.2
0	65	4.9	147	11.2
1	611	46.5	758	57.6
2	557	42.4	1315	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	101	7.7	101	7.7
0	149	11.3	250	19.0
1	632	48.1	882	67.1
2	433	32.9	1315	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	240	18.3	240	18.3
0	362	27.5	602	45.8
1	233	17.7	835	63.5
2	480	36.5	1315	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	217	16.5	217	16.5
0	283	21.5	500	38.0
1	346	26.3	846	64.3
2	469	35.7	1315	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 17

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	117	8.7	117	8.7
0	244	18.2	361	26.9
1	375	27.9	736	54.8
2	606	45.2	1342	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	106	7.9	106	7.9
0	136	10.1	242	18.0
1	602	44.9	844	62.9
2	498	37.1	1342	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	131	9.8	131	9.8
0	138	10.3	269	20.0
1	616	45.9	885	65.9
2	457	34.1	1342	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	214	15.9	214	15.9
0	629	46.9	843	62.8
1	345	25.7	1188	88.5
2	154	11.5	1342	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	157	11.7	157	11.7
0	804	59.9	961	71.6
1	264	19.7	1225	91.3
2	117	8.7	1342	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	184	13.7	184	13.7
0	722	53.8	9063	67.5
1	436	32.5	1342	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	294	21.9	294	21.9
0	584	43.5	878	65.4
1	304	22.7	1182	88.1
2	160	11.9	1342	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	277	20.6	277	20.6
0	564	42.0	841	62.7
1	406	30.3	1247	92.9
2	95	7.1	1342	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 18

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	65	4.8	65	4.8
0	171	12.5	236	17.3
1	236	17.3	472	34.5
2	895	65.5	1367	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	116	8.5	116	8.5
0	203	14.9	319	23.3
1	375	27.4	694	50.8
2	673	49.2	1367	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	161	11.8	161	11.8
0	196	14.3	357	26.1
1	348	25.5	705	51.6
2	662	48.4	1367	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	208	15.2	208	15.2
0	388	28.4	596	43.6
1	484	35.4	108	79.0
2	258	18.9	01338	97.9
3	29	2.1	1367	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	122	8.9	122	8.9
0	533	39.0	655	47.9
1	561	41.0	1216	89.0
2	151	11.0	1367	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	192	14.0	192	14.0
0	230	16.8	422	30.9
1	587	42.9	1009	73.8
2	358	26.2	1367	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	206	15.1	206	15.1
0	576	42.1	782	57.2
1	434	31.7	1216	89.0
2	151	11.0	1367	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	207	15.1	207	15.1
0	588	43.0	795	58.2
1	276	20.2	1071	78.3
2	296	21.7	1367	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Form 19

Constructed-Response 1

Raw14	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	62	5.0	62	5.0
0	81	6.5	143	11.4
1	115	9.2	258	20.6
2	254	20.3	512	40.9
3	740	59.1	1252	100.0

Constructed-Response 2

Raw15	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	74	5.9	74	5.9
0	63	5.0	137	10.9
1	251	20.5	394	31.5
2	858	68.5	1252	100.0

Constructed-Response 3

Raw16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	169	13.5	169	13.5
0	175	14.0	344	27.5
1	206	16.5	550	43.9
2	702	56.1	1252	100.0

Constructed-Response 4

Raw30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	192	15.3	192	15.3
0	260	20.8	452	36.1
1	411	32.8	863	68.9
2	296	23.6	1159	92.6
3	93	7.4	1252	100.0

Table 25 (cont). HSPT in Science Pilot
Frequency Distribution for Constructed-Response Item Responses

Constructed-Response 5

Raw31	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	195	15.6	195	15.6
0	171	13.7	366	29.2
1	280	22.4	646	51.6
2	606	48.4	1252	100.0

Constructed-Response 6

Raw32	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	309	24.7	309	24.7
0	254	20.3	563	45.0
1	517	41.3	1080	86.3
2	172	13.7	1252	100.0

Constructed-Response 7

Raw46	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	220	17.6	220	17.6
0	413	33.0	633	50.6
1	328	26.2	961	76.8
2	291	23.2	1252	100.0

Constructed-Response 8

Raw50	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	247	19.7	247	19.7
0	493	39.4	740	59.1
1	134	10.7	874	69.8
2	378	30.2	1252	100.0

Table 26. HSPT in Science Pilot
Group Descriptive Statistics

	White			African-American			Female			Male		
Form	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
12	36.44	9.76	1032	23.05	9.29	129	34.17	9.84	664	34.8	11.54	692
13	32.95	9.89	1018	21.6	8.22	144	30.73	9.98	680	31.69	10.98	646
14	34.05	11.03	1026	22.57	9.48	89	32.40	10.53	628	33.64	12.23	654
15	34.49	9.93	987	22.6	9.11	43	33.02	9.65	591	34.28	11.06	579
16	35.82	10.01	1060	29.35	9.22	83	34.62	9.56	663	35.40	11.03	633
17	32.17	10.34	1031	24.00	9.47	142	31.63	10.05	658	30.48	11.11	648
18	35.77	10.07	1067	34.23	10.25	115	33.59	9.95	651	34.84	11.43	678
19	35.96	10.77	991	21.58	8.75	86	34.07	10.79	614	35.29	11.66	589

Table 27. HSPT in Science Pilot
DIF Statistics (Standardized Mean Differences: SMDs) for Gender and Ethnic Groups

Gender

Form	# of Items	# of Males	# of Females	DIF Against Males SMD≥.20	DIF Against Males .19≥SMD≥.10		DIF Against Females SMD≤-.20	DIF Against Females -.19≤SMD≤-.10
12	50	692	664	1	1	(1)*	0	2
13	50	646	680	1	6	(3)	0	5
14	50	654	628	0	5	(3)	0	2
15	50	579	591	1	5	(2)	0	5
16	50	633	663	1	3	(0)	0	5
17	50	648	658	0	6	(2)	1	6
18	50	678	651	1	2	(1)	0	5
19	50	589	614	1	6	(3)	1	3

Ethnicity

Form	# of Items	# of Whites	# of African-Americans	DIF Against Whites SMD≥.20	DIF Against Whites .19≥SMD≥.10		DIF Against African-Americans SMD≤-.20	DIF Against African-Americans -.19≤SMD≤-.10
12	50	1032	129	0	5	(2)*	2	3
13	50	1018	144	0	4	(1)	0	5
14	50	1026	89	0	6	(4)	2	6
15	50	987	43	1	4	(2)	0	8
16	50	1060	83	0	7	(1)	1	4
17	50	1031	142	0	4	(1)	0	5
18	50	1067	115	0	4	(0)	2	0
19	50	991	86	0	5	(0)	2	1

* Absolute value of the difference in total "practically significant" DIF across the two groups of a comparison. Total DIF for each group is twice the number of items with |SMD|≥.20 plus the number of items with .10≤|SMD|≤.19.

Appendix D

Science
Student Survey

Directions: *Listed below are statements about activities that often take place in mathematics classes. The Michigan Department of Education is interested in finding out how often these activities have been a part of <u>your</u> school experience by the end of tenth grade.*

Please read each question carefully and answer it the BEST that you can. For each question, darken one circle on your answer sheet labeled Session 1 *to indicate your response using the scale below.*

Scale:	A	B	C	D
	Never	Very Little	Some	A lot

Sample Item:

By the end of tenth grade, how often did your school experience include:

A: using trigonometric ratios to solve problems involving sine and cosine?

By the end of tenth grade, how often did your school experience include:

1. discussing current scientific events from newspapers, magazines, or television?
2. taking short answer tests in science?
3. taking essay tests in science?
4. taking cluster-type science tests (problems followed by a few multiple-choice questions)?
5. following procedures to complete a laboratory experiment?
6. making diagrams to explain your thinking?
7. using charts, graphs, tables, or diagrams to answer questions?
8. writing explanations about what you observed and why it happened?
9. discussing contributions to science from cultures and individuals of diverse backgrounds?
10. making judgments and explaining your reasons about how to best solve a real-life science problem (e.g., pollution)?
11. using information you have collected to make charts, graphs, or tables?
12. making predictions about things that happen and how they are related?
13. designing your own experiment or investigation?
14. critiquing the results of an experiment or investigation?
15. relating what you have learned in science class to the real world?
16. giving the reasons behind an incorrect hypothesis?

Listed below are science topics often taught by the end of tenth grade. Next to each topic is the description. Please read them carefully and then estimate how often you studied each topic by the end of tenth grade. For each topic, darken one circle on your answer sheet to indicate your response using the same scale you used for the questions above.

TOPIC		DESCRIPTION
17.	Cells	Cell structures, kinds of cells, how cells grow, develop, and reproduce
18.	Organization of Living Things	Classifying organisms, life cycle, photosynthesis, and adaptations
19.	Heredity	How genetic traits are passed on, sexual and asexual reproduction, DNA replication
20.	Evolution	Tracing origin and development of species, adaptations, natural selection, changes in living things
21.	Ecosystems	Ecological relationships, flow of energy in ecosystems, factors regulating population size, natural cycles, effect of humans on ecosystems
22.	Geosphere	Surface features of earth, map study, geological history, plate tectonics, rock cycle, conservation practices
23.	Hydrosphere	Forms of water, river water flow, pollution in atmosphere, climate, water quality
24.	Atmosphere and Weather	Water cycle, patterns of air movement, weather predictions, climate changes, impact on humans
25.	Space Science	Formation of solar system, rotation and revolution of planets, seasons, sun, instruments used in space study (telescopes, etc.)
26.	Matter and Energy	Elements, compounds, mixtures, atoms, density, electricity, magnetic fields, circuits, heat transfer
27.	Changes in Matter	Chemical, physical, nuclear, conservation of mass, energy transformations, chemical bonds
28.	Motion of Objects	Speed, direction, changes in direction, action-reaction, force and motion, magnetic forces, potential and kinetic energy
29.	Waves and Vibrations	Properties of sound waves, light, colors, spectrum, kinds of waves, electromagnetic spectrum, recording devices, transfer of energy by waves

Thank you very much!

Table 30. Student Survey Response Means

By the end of tenth grade, how often did your school experience include:

Statement #	Statement	Mean
13*	designing your own experiment or investigation?	.99
9*	discussing contributions to science from cultures and individuals of diverse backgrounds?	1.04
23*	hydrosphere?	1.39
22*	geosphere?	1.39
3*	taking essay tests in science?	1.41
10*	making judgments and explaining your reasons about how best to solve a real-life science problem (e.g., pollution)?	1.42
1	discussing current scientific events from newspapers, magazines, or television?	1.43
6	making diagrams to explain your thinking?	1.44
14	critiquing the results of an experiment or investigation?	1.49
25*	space science?	1.50
15*	relating what you have learned in science class to the real world?	1.56
16*	giving the reasons behind an incorrect hypothesis?	1.62
29*	waves and vibrations?	1.63
24	atmosphere and weather?	1.64
21*	ecosystems?	1.65
20	evolution?	1.78
12	making predictions about things that happen and how they are related?	1.84
28	motion of objects?	1.88
11	using information you have collected to make charts, graphs, or tables?	1.90

Table 30 (cont). Student Survey Response Means

Statement #	Statement	Mean
7	using charts, graphs, tables, or diagrams to answer questions?	1.96
2	taking short answer tests in science?	1.99
8	writing explanations about what you observed and why it happened?	2.12
27	changes in matter?	2.12
19	heredity?	2.23
26	matter and energy?	2.25
4	taking cluster-type science tests (problems followed by a few multiple-choice questions)?	2.27
5	following procedures to complete a laboratory experiment?	2.34
18	organization of living things?	2.39
17	cells?	2.59

*** - more than 10% of students responded "never".**

Table 31. Teacher Survey - Science
Statements with ≥ 50% Schools Responding NSI

(N = 244)

Statement		% of Schools Responding NSI[5]	% of Schools Responding NT[6]
71	Explain how sound recording and reproducing devices work. PWV13. Parts of sound recording and reproducing devices, including: needle, amplifier, speaker, microphone, laser disk reader	60%	28%
52	Analyze properties of common household and agricultural materials in terms of risk/benefit balance PME19. Risk/benefit analysis	56%	28%

Table 32. Teacher Survey - Science
Statements with 0% Schools Responding NT

1. Classify cell and organisms on the basis of organelle and/or cell types

   LC5. Cell parts used for classification: organelle, nucleus, cell wall, cell membrane

6. Explain how cells use food to grow, and how materials move into and out of cells

   LC10. Words describing how materials pass in and out of cells: osmosis, diffusion

[5] NSI = Not Sufficient Instruction
[6] NT = Not Taught

MICHIGAN HIGH SCHOOL PROFICIENCY TEST IN SCIENCE

Tryout and Pilot Technical Report Development Team

(alphabetically)

Jane K. Faulds

Catherine B. Smith

Jean W. Yan

Correspondence concerning this report should be addressed to:

Jean Yan or Catherine Smith
MEAP Office
Michigan Department of Education
P.O. Box 30008
Lansing, MI 48909
(517) 373-8393 (o)
(517) 335-1186 (fax)
yanj@state.mi.us
smithcb@state.mi.us

This development team wishes to thank the following people for their time and expertise in reviewing this document and providing suggestions and comments: Drs. Burt Voss and Nancy Shiffler. CTB/McGraw-Hill as contractor for the development phase of the HSPT in Mathematics, Reading and Science, provided data and most statistical analyses used in this report.

Anastasia M. Gormely and Yolanda Y. Stephens
provided excellent support services for this project.